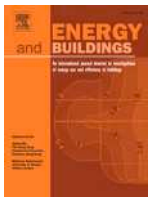
Deploying deep reinforcement learning for low-level HVAC control in multi-zone buildings: A comparative study with ASHRAE G36 sequences

Energy & Buildings 348 (2025) 116456

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild



Deploying deep reinforcement learning for low-level HVAC control in multi-zone buildings: A comparative study with ASHRAE G36 sequences

Sabrina Savino [a], Giuseppe Razzano [a], Michele Pagone [b], Carlo Novara [b], Alfonso Capozzoli [a,*]

[a] Department of Energy (DENERG), TEBE Research Group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy
[b] Department of Electronics and Telecommunications (DET), Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy

ARTICLE INFO

Keywords:
AHU low-level control
Multi-zone building
Deep reinforcement learning
ASHRAE guideline 36 (G36)
Modelica

ABSTRACT

This paper proposes a methodology for optimizing HVAC control in multi-zone buildings using Deep Reinforcement Learning. The study focuses on optimizing the central AHU system by controlling all low-level components within both the air and water loops, addressing the complex dynamics of multi-zone interactions. The case study is based on a building within the Politecnico di Torino campus. Modelica-based simulations are used to model both the HVAC system and building dynamics, allowing the integration and evaluation of the ASHRAE G36 control standard as a benchmark. Two DRL strategies are developed and evaluated, Zone-Aware and Zone-Integrated, under both winter and summer conditions, with the goal of improving energy efficiency, indoor temperature control, and indoor CO_2 concentration, under varying occupancy profiles. The results reveal that both DRL strategies outperform the G36 baseline in terms of energy savings (up to 17 %), indoor temperature violations, and CO_2 concentration. Additionally, DRL controllers demonstrate strong generalizability and adapt seamlessly to unseen occupancy profiles without manual tuning. This research highlights the potential of DRL to provide scalable, adaptive, and energy-efficient HVAC control solutions for multi-zone buildings.

1. Introduction

The building sector is responsible for more than a third of global energy consumption and emissions [1]. Among building systems, heating, ventilation and air conditioning (HVAC) systems account for almost 50 % of the total energy in office and residential buildings [2]. Ventilation plays a crucial role in maintaining thermal comfort and indoor air quality (IAQ) by providing and distributing conditioned outdoor air across different building zones. However, this process requires significant energy input due to the operation of complex networks of fans and ducts. Although proper ventilation prevents air stagnation and mitigates indoor pollutant accumulation, excessive airflow leads to unnecessary energy consumption, whereas inadequate ventilation compromises air quality and may pose health risks [3,4]. The balance of energy efficiency and occupant well-being remains a critical challenge in the management and control of the HVAC system.

This challenge is further amplified in multi-zone HVAC systems, where nonlinear airflow dynamics and inter-zone interactions make the control problem highly complex. Due to shared ductwork and ventilation networks, adjustments in one zone can unintentionally affect airflow and temperature conditions in adjacent spaces [5,6]. Furthermore, determining a single optimal supply air temperature for all zones is challenging, as thermal loads and occupant preferences vary [7]. These complexities require adaptive control strategies capable of dynamically managing ventilation to optimize both energy use and occupant comfort.

Traditional HVAC systems rely on rule-based hierarchical control frameworks, where a supervisory controller adjusts setpoints for lower-level controllers that regulate actuators (e.g., valves, dampers, and fans). ASHRAE Guideline 36 (G36) [8] provides a structured approach for optimizing Variable Air Volume (VAV) systems in multi-zone buildings. Incorporates supply air temperature reset, duct static pressure reset, and zone airflow control strategies to improve energy efficiency. Simulation-based studies indicate that G36 can achieve 31 % HVAC energy savings in medium-sized commercial buildings compared to conventional rule-based methods [9]. Despite these advantages, rule-based controllers remain static and struggle to adapt to dynamic environmental conditions. Additionally, tuning Proportional-Integral-Derivative (PID) controllers in multi-zone systems is labor-intensive, requiring expert calibration to maintain optimal performance under varying occupancy patterns, weather conditions, and system loads.

* Corresponding author.
E-mail address: alfonso.capozzoli@polito.it (A. Capozzoli).

https://doi.org/10.1016/j.enbuild.2025.116456
Received 1 July 2025; Received in revised form 26 August 2025; Accepted 14 September 2025
Available online 16 September 2025

To address these limitations, Model Predictive Control (MPC) and Deep Reinforcement Learning (DRL) have emerged as promising alternatives for HVAC optimization. MPC uses predictive modeling to optimize control actions over a given time horizon [10]. It has been successfully implemented in single-zone systems [11,12] and multizone buildings with independent HVAC units per zone [13]. However, its application in fully interconnected multi-zone HVAC systems is challenging due to the need for detailed, calibrated models that capture complex inter-zone airflow interactions. This requires high computational resources and continuous model updates, making large-scale deployment cumbersome [14], especially for multi-agent network systems [15].

In contrast, DRL provides a model-free, data-driven approach that learns optimal control policies through direct system interaction. Unlike MPC, DRL does not require an explicit system model, instead leveraging empirical observations and reward signals to adapt control strategies in real time. This makes DRL particularly well-suited for scenarios where system modeling is complex or impractical and in dynamic HVAC environments with changing occupancy, weather conditions, and equipment performance. However, training DRL models presents challenges, as it requires extensive data and must carefully balance exploration and exploitation to prevent suboptimal or unsafe actions [16]. To mitigate the potential risks associated with unsafe actions, advanced co-simulation environments have been developed, enabling realistic testing of control strategies before deployment [17,18]. Integrating multiple simulation platforms within a co-simulation framework effectively captures the dynamic interactions between HVAC components, thus improving the evaluation of complex control strategies [19]. For instance, integrating Modelica models as Functional Mock-up Units (FMUs) within co-simulation frameworks allows for high-fidelity HVAC modeling and seamless tool interoperability [20]. Despite these challenges, DRL and co-simulation environments complement each other by combining adaptive learning with detailed simulation-based testing, enabling safer and more effective deployment. This synergy exhibits strong potential for solving complex multi-objective optimization problems, making DRL a promising approach for HVAC control.

This paper explores the role of DRL in multi-zone HVAC control, examining its state-of-the-art applications, limitations, and benchmarking against existing standards like ASHRAE G36.

1.1. Related works and contributions

HVAC system control operates at two primary levels: low-level control, which directly manages actuators (e.g., dampers, fans, valves), and supervisory control, which sets target values (setpoints) that guide low-level controllers, typically PID-based. Research efforts have focused on improving supervisory control through advanced optimization techniques, mainly by dynamically adjusting temperature and pressure setpoints to improve energy efficiency and occupant comfort [21–26].

Despite these advancements, a key study [25] directly compares the performance of advanced controllers, including DRL, against the G36 control strategy in a multi-zone setting. Their findings suggest that DRL at the supervisory level does not provide a clear advantage over G36 in terms of energy savings and thermal comfort. Both methods yield comparable results, indicating that optimizing set points alone may not be sufficient for superior HVAC performance.

One fundamental limitation of supervisory control is its reliance on predefined or dynamically adjusted setpoints, which do not always translate to optimal low-level control actions. Dynamic setpoint changes often require the corresponding PID tuning to ensure a proper system response. This challenge is further intensified by the non-linear and time-varying nature of building and HVAC systems, which require frequent adjustments to PID parameters in response to weather fluctuations, occupancy shifts, and equipment performance variations. Furthermore, even with robust control-oriented building models, supervisory control remains a complex nonlinear optimization problem where effectiveness depends on solver tuning and initialization [27].

To address these challenges, several studies have explored hybrid control strategies that integrate supervisory and low-level control actions [28,29]. However, these approaches typically optimize a multi-objective function while focusing only on a subset of key factors–such as energy consumption, zone temperature violations, or indoor air quality (e.g., CO_2 levels)–without achieving full integration with low-level control.

Recent efforts have begun addressing this gap. For instance, the authors in [30] optimizes energy consumption, thermal comfort, and indoor air quality (CO_2 and humidity levels) by jointly controlling the supply temperature setpoint, fan speed, outdoor damper position, and ventilation air operation in a single-zone office building. In the multi-zone context, [31] introduces an imitation-interaction learning method to directly control low-level components, such as fan speed and VAV dampers, improving ventilation efficiency to enhance energy savings.

However, comprehensive low-level control optimization for full multi-zone HVAC systems remains an underexplored area, and the relevant literature is rather limited. An alternative, yet unexplored approach, is to optimize only the AHU-level control, while maintaining existing operational settings and control strategies for VAV controllers. Since each VAV typically serves a single zone, this strategy reduces the number of control actions required, offering a scalable and efficient path to multi-zone HVAC optimization without the complexity of modifying individual VAVs.

Furthermore, few studies employ high-fidelity simulation models to accurately represent the dynamics of the HVAC system and evaluate advanced control strategies. OpenModelica [32], an equation-based simulation environment, provides a more detailed representation of HVAC systems, enabling seamless integration of both mechanical and control components in a single framework. Moreover, OpenModelica enables a straightforward implementation of the ASHRAE Guideline 36 logic, which is challenging to achieve in other simulation environments such as EnergyPlus [33]. This capability enables a direct, high-fidelity benchmarking of advanced control strategies [34], including DRL, against widely recognized standards. However, only a limited number of studies (see, e.g., [25,35,36]) adopt the Modelica language [37] as a high-fidelity plant to test advanced control law in multi-zone HVAC systems, leaving a significant gap in assessing an adequate methodology to compare the performance of advanced strategies with respect to those of G36.

This study aims to bridge these gaps by developing DRL-based control strategies for a multi-zone Air Handling Unit (AHU) system, evaluated in a Modelica-based high-fidelity environment. The key contributions are the following:

- Centralized AHU-based Multi-Zone Control: instead of directly controlling individual VAV actuators, the proposed methods act on low-level components within the centralised AHU, allowing the VAVs to continue operating under the G36 control logic. This setup enables an evaluation of centralized control performance and its scalability to larger multi-zone systems.
- Direct low-level Control of AHU actuators: unlike supervisory approaches that adjust setpoints, this study directly acts with a DRL-based controller on AHU components (e.g., dampers, valves, and fan speed) to optimize energy consumption, temperature deviations across zones, and indoor CO_2 levels.
- Development and Comparison of two DRL strategies:
  - The *Zone-Aware* DRL strategy leverages full observability of each zone's temperature to make zone-specific control decisions.
  - The *Zone-Integrated* DRL strategy replaces individual zone temperature measurements with two engineered variables representing the aggregated thermal violations above the upper bound and below the lower bound of the desired indoor temperature band. This approach mimics G36's voting mechanism and simplifies the observation space.

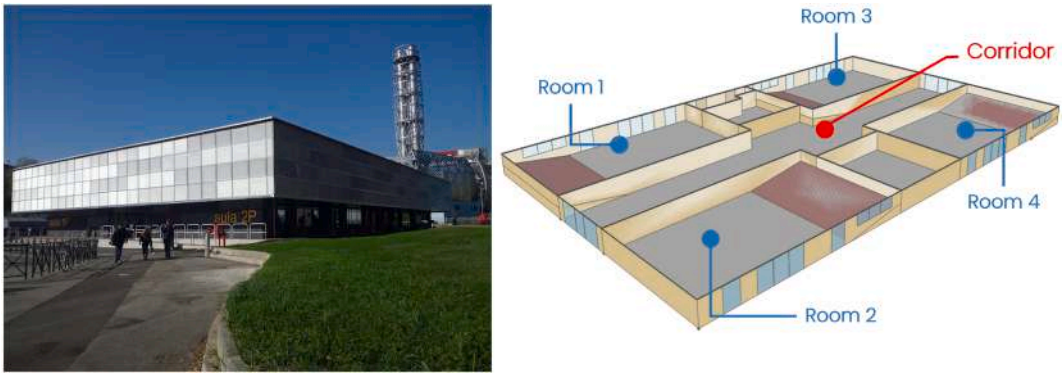


Fig. 1. Educational building of Politecnico di Torino campus [38].

- Adaptability to Diverse Occupancy Profiles: the proposed low-level DRL controllers acting directly on AHU actuators were tested under different occupancy patterns. The results demonstrate robust performance and adaptability without requiring additional fine-tuning of the learned policy.
- High-fidelity Benchmarking with Modelica: by leveraging Modelica's advanced modeling capabilities, the proposed DRL strategies are compared directly with the G36 standard to assess its effectiveness in improving energy efficiency, indoor temperature control, and CO_2 levels.

2. Case study: building and AHU system configurations

The selected case study is an educational building situated within the Politecnico di Torino campus. The building consists of five distinct thermal zones: four classrooms and a central corridor, as depicted in Fig. 1. Detailed thermo-physical and geometrical properties are summarized in Table 1.

The building operates according to a fixed occupancy schedule, with active occupancy from 08:00 to 19:45, Monday through Friday. During weekends, specifically Saturdays and Sundays, the building remains unoccupied. The setpoints of indoor air temperature vary depending on occupancy and operational mode. During occupied hours, the setpoint temperature (T_{sp}) is maintained at 21 °C for heating and 26 °C for cooling. Outside occupied periods, setpoint temperatures are relaxed to 15 °C for heating and increased to 30 °C for cooling, allowing energy savings when spaces are unoccupied.

While the building geometry and thermo-physical properties of the envelope are derived from the actual case study, the HVAC configuration in the simulation does not correspond to the real system, which is equipped with five independent rooftop units (RTUs). To enable coordinated multi-zone control strategies, the model instead assumes a centralized AHU with local VAV terminal boxes. This setup also facilitates comparison with standardized baselines such as ASHRAE Guideline 36 and accounts for both heating and cooling operation. A scheme of the HVAC configuration considered in this study is illustrated in Fig. 2. The system comprises a central air handling unit (AHU) serving all five thermal zones, equipped with an economizer system for efficient use of outdoor air, integrated heating and cooling coils, a supply fan, and five Variable Air Volume (VAV) boxes. Airflow regulation is managed through dedicated VAV dampers. Cooling is provided by a central chiller, while heating is primarily delivered by a heat pump. Zone-level heating adjustments are handled by electric resistance reheat coils integrated into the variable air volume (VAV) terminals.

The AHU operates through six primary control signals, which vary based on the current operation mode, either heating or cooling. Regardless of the mode, an economizer damper control signal modulates outdoor and return air damper positions, thus adjusting the intake of outdoor air ($\dot{m}_{out}$) and recirculated return air ($\dot{m}_{ret}$). This modulation ensures that the mixed air temperature (T_{mix}) within the AHU is properly maintained in response to variations in outdoor air temperature (T_{out}) and return air temperature (T_{ret}). Additionally, fan speed is adjusted through a fan RPM control signal, determining the total supply air mass flow rate ($\dot{m}_{tot}$) delivered to the zones, thus balancing ventilation and comfort requirements.

During the cooling season, the system maintains a fixed chilled water temperature at 6 °C. The cooling coil valve control signal continuously modulates the chilled water mass flow rate ($\dot{m}_{water}$) through the cooling coil, thereby precisely regulating cooling capacity. Conditioned air leaving the AHU is distributed to the individual zones, where VAV dampers adjust the discharge airflow rates ($\dot{m}_{dis}$) to each zone, maintaining indoor temperature conditions according to the defined cooling setpoints.

During the heating season, hot water is supplied to support zone-level heating, which is delivered via post-heating coils integrated into each of the VAV boxes. These post-heating coils, which operate using electrical resistance heating, provide supplementary heating directly at the zone level. The hot water mass flow rates through these coils are regulated by a dedicated post-heating coil valve control signal, ensuring accurate and responsive temperature control within each zone. Simultaneously, the VAV dampers modulate the supply air mass flow rates ($\dot{m}_{dis}$), allowing the HVAC system to consistently achieve the specified heating setpoints in each thermal zone.

Table 1
Description of building features.

Building feature	Value	Unit
N° of thermal zones	5	[-]
Conditioned Room Volume	948.50	[m^3]
Conditioned Corridor volume	1361.09	[m^3]
Transparent/opaque envelope vertical surface ratio	0.65	[-]
Opaque envelope vertical surface	197.2	[m^2]
U-Value Exterior Wall	0.15	[W/m^2K]
U-Value Interior Wall	0.21	[W/m^2K]
U-Value Floor	0.57	[W/m^2K]
U-Value Roof	0.15	[W/m^2K]
U-Value Window	1.6	[W/m^2K]

3. Methodology

This section outlines the methodology employed in the study, detailing the modeling of the building and HVAC system, the co-simulation framework that enables interaction with external controllers, and the Deep Reinforcement Learning (DRL) control strategies. A schematic overview of the methodology is illustrated in Fig. 3.

3.1. Building and HVAC system modeling

The simulation environment models both the building and the Air Handling Unit (AHU) system, along with its associated components,

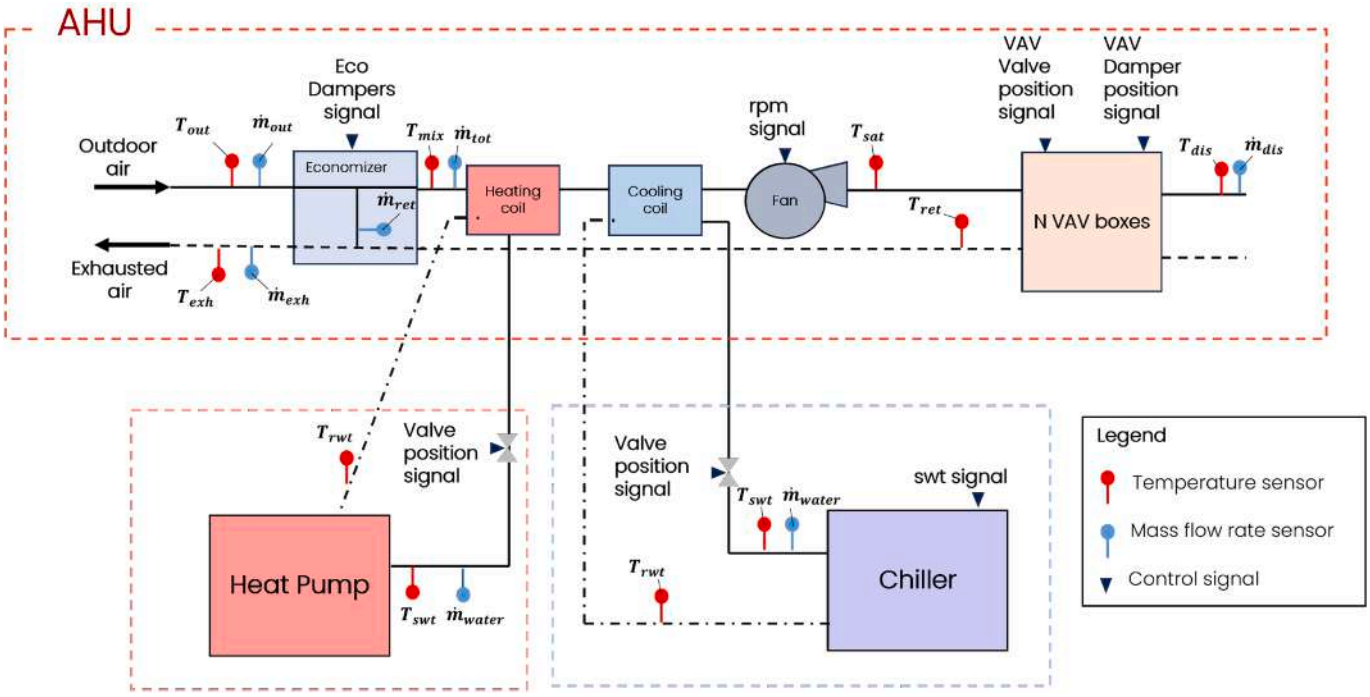


Fig. 2. Schematic representation of the AHU system [39].

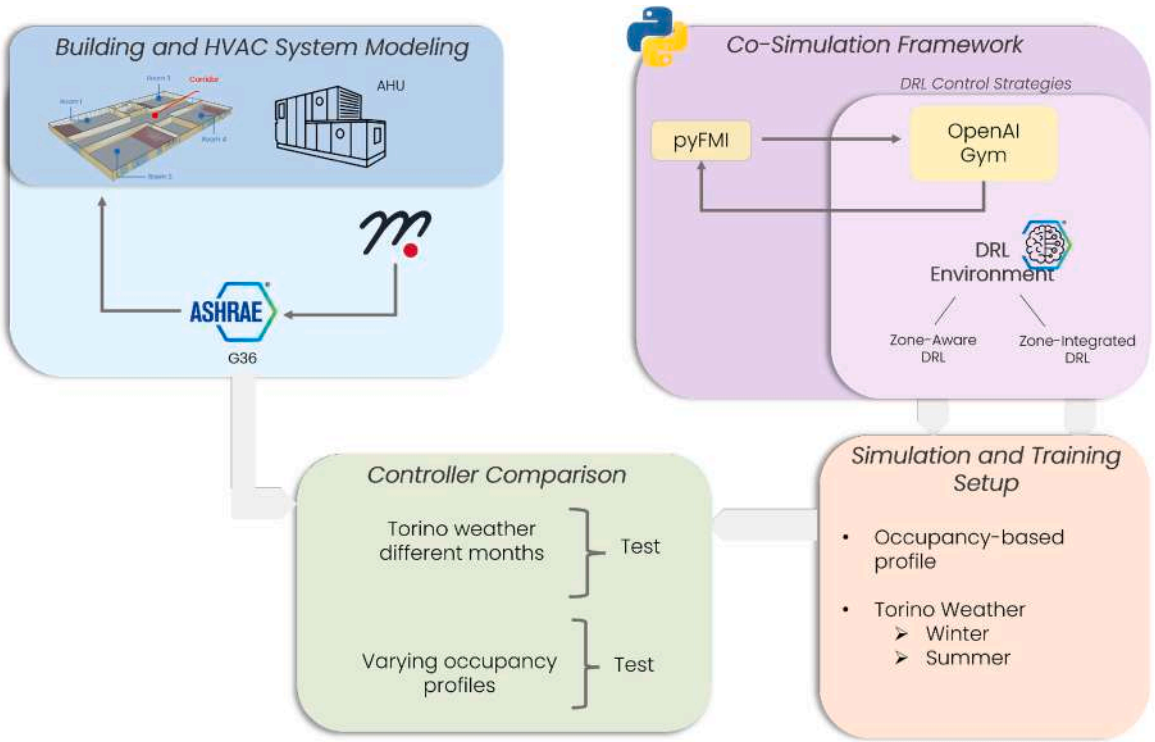


Fig. 3. Methodological framework.

using the Modelica language within the OpenModelica open source platform [32], in combination with the Buildings Library [40]. In this study, we employed a reduced-order formulation that represents the opaque envelope components, such as exterior walls and roof, through a network of thermal resistances and capacitances. Compared to simpler lumped approaches, the FourElements model in the Buildings Library [40] provides a finer resolution of building dynamics by distinguishing between different envelope elements, thereby capturing their specific heat transfer characteristics while preserving computational efficiency. Users can adjust the complexity of the model by selecting the number of wall elements and the level of spatial discretization based on the thermal properties of the walls and external excitations. In this study, the stratigraphy of the building envelope was derived from available architectural documentation to ensure a realistic representation of thermal behavior.

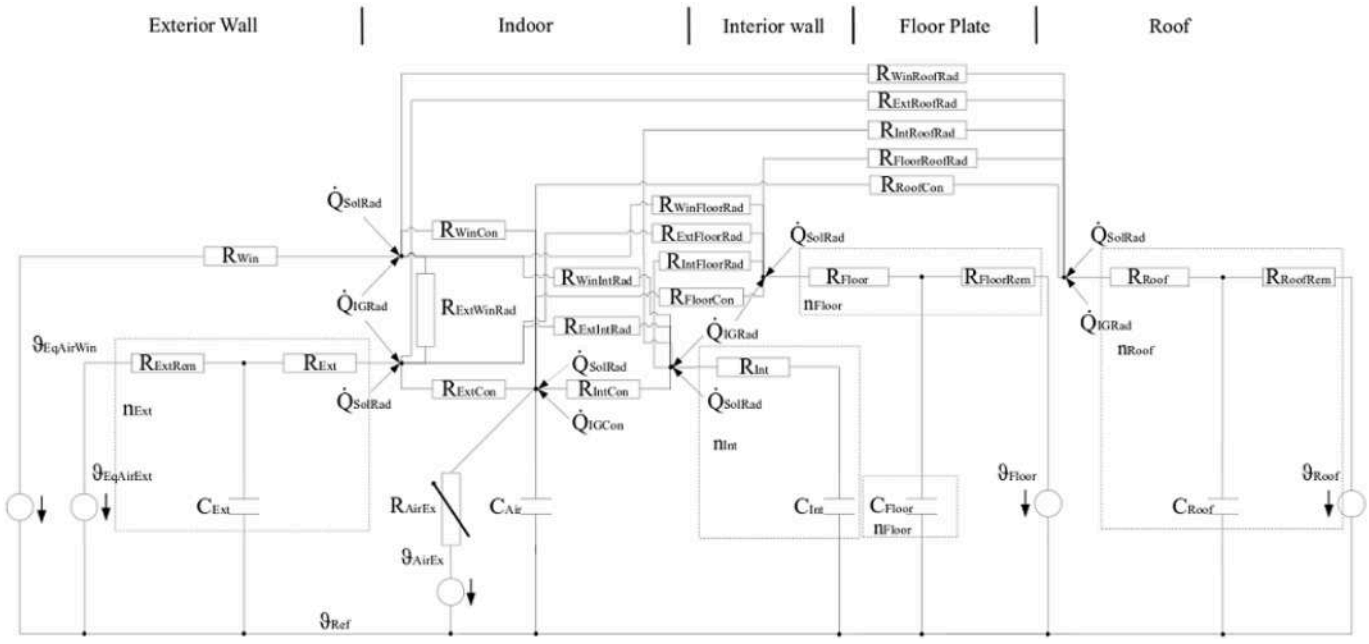


Fig. 4. Lumped-parameter representation of thermal dynamics in a single zone.

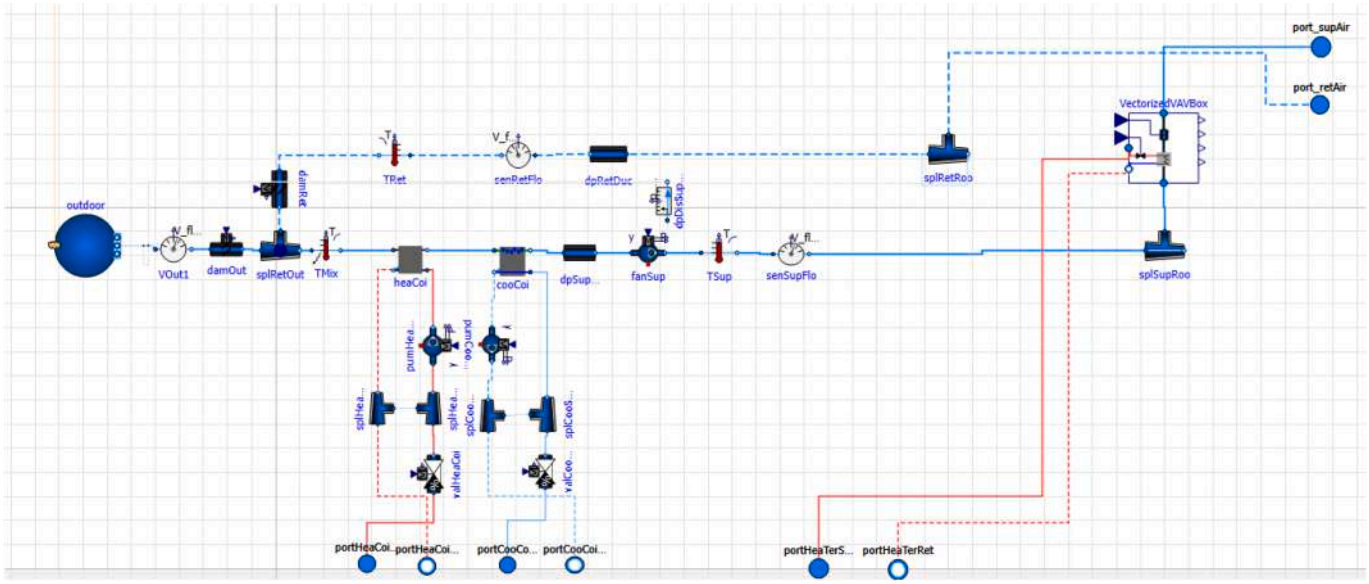


Fig. 5. AHU model in OpenModelica.

In the proposed configuration, five thermal zones are modeled and interconnected. The RC circuit representation of a single thermal zone is shown in Fig. 4.

The HVAC system introduced in Section 2 is implemented in OpenModelica, allowing for high-fidelity simulation of its dynamic behavior and control strategies. The implementation includes detailed representations of key AHU components–such as heating and cooling coils, economizer dampers, supply fans, and VAV boxes with reheat coils–capturing both thermal and airflow dynamics. Components and actuators are modeled to reflect realistic operational responses (Fig. 5). To establish a performance benchmark, the ASHRAE Guideline 36 standard is implemented within the Buildings Library to model the operation of the AHU and the VAVs' components. This provides a standardized reference framework, ensuring that advanced control strategies, such as DRL-based approaches, can be effectively evaluated.

3.2. *Co-simulation framework*

To ensure seamless integration between the building model, HVAC system, and DRL control algorithms, a co-simulation approach is adopted. The Modelica-based model is exported as a Functional Mock-up Unit (FMU) using the Functional Mock-up Interface (FMI) standard [20]. This facilitates interaction with external software tools while preserving the integrity of the physical system model.

Python is employed as the central coordination framework, managing simulation execution, data exchange, and real-time interactions between system components. The pyfmi package [41] is used to facilitate efficient communication between the FMU, encapsulating the building thermal model, the AHU system, and the control algorithm, responsible for generating actuator control signals. This framework enables highly flexible system integration, allowing for dynamic interactions between

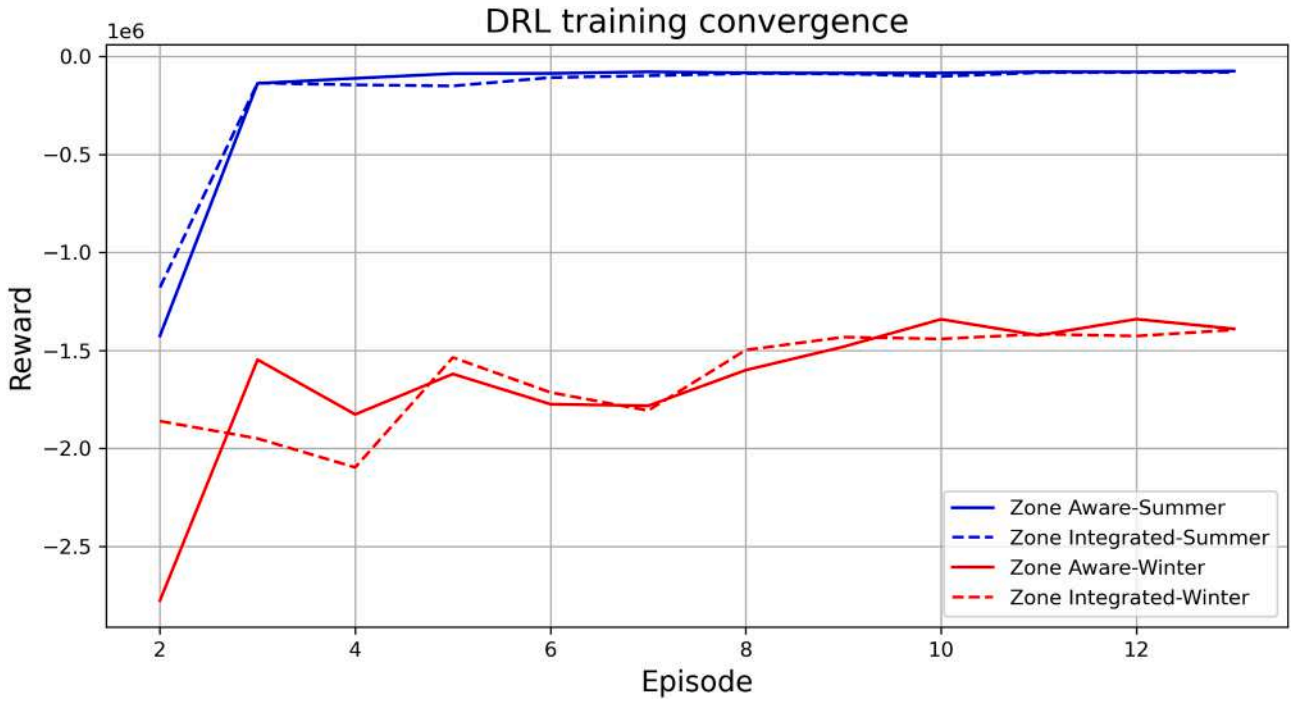


Fig. 6. Convergence of DRL agents during training. In summer (blue), the cumulative penalty rapidly stabilizes within the first three episodes, while in winter (red), it converges more gradually and plateaus after about nine episodes. The first episode excluded as it corresponds to the warm-up phase without policy updates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

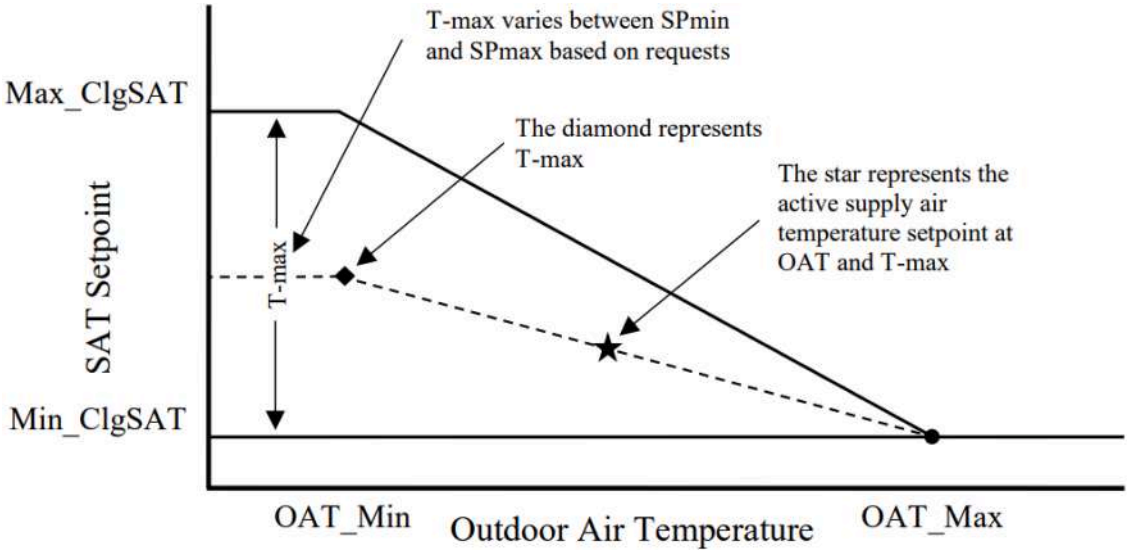


Fig. 7. Supply air temperature reset diagram [8].

building physics, HVAC operation, and advanced control methodologies.

3.3. DRL control strategies and training setup

Deep Reinforcement Learning (DRL) controllers are integrated into the simulation framework using the OpenAI Gym environment [42], enabling an adaptive control strategy that continuously interacts with the environment, learns from system dynamics, and iteratively refines its policy to optimize HVAC operations.

Two DRL control strategies are implemented: Zone-Aware DRL and Zone-Integrated DRL, differing in their observation spaces. Further details are provided in Section 4.2. Both agents are trained using the Soft Actor-Critic (SAC) algorithm implemented in Stable-Baseline3 [43]. The main hyperparameters are: multilayer perceptron policy with four hidden layers of 64 units (ReLU activations), discount factor $\gamma = 0.99$, Adam optimizer with learning rate $lr = 10^{-3}$, batch size 128, replay buffer size of 10^6, entropy coefficient set to auto, soft target update coefficient $\tau = 0.005$, and training starting after 10,000 warm-up steps.

To evaluate the proposed methodology under realistic conditions of the case study, climate data from Turin, Italy, were sourced from [44]. The DRL training phase is conducted separately for each season, spanning one month each. Specifically, training for the heating season takes place from January 1, to January 31, while training for the cooling season occurs from July 1, to July 31, to ensure season-specific adaptation. The use of separate DRL training for the different seasons is motivated by

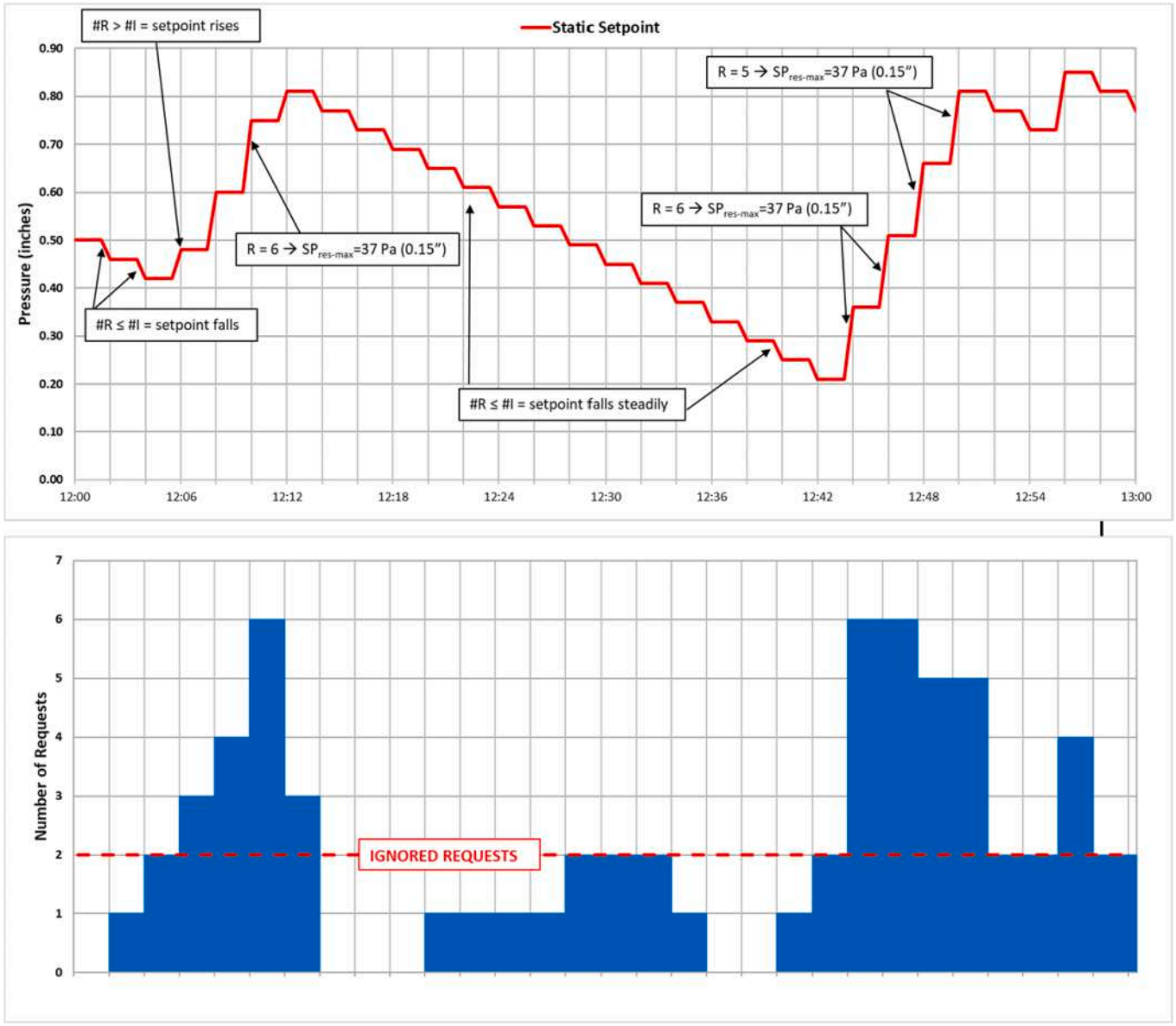


Fig. 8. Static pressure reset diagram [8].

differences in system configuration given by the presence of post-heating in the VAV boxes during the heating season. Each training episode corresponds to a full month of simulation, and the agents were trained for 13 episodes. The learning curves in Fig. 6 show the cumulative reward per episode for both DRL strategies in summer and winter conditions.

3.4. Controllers comparison

Following the training phase, the DRL controllers are deployed for an extended seasonal period to assess their performance in realistic building operation scenarios. The cooling season deployment runs from June 20, to August 20, whereas the heating season deployment extends from January 1, to February 28. This deployment phase enables a comprehensive evaluation of the controllers' ability to regulate indoor conditions dynamically across varying climatic scenarios.

To assess the robustness and generalization of the DRL controllers under unseen conditions, additional test cases with varying occupancy profiles are evaluated (see Section 5 for further details). The DRL policies are deployed statically, without retraining or fine-tuning, to examine their ability to maintaining energy efficiency, thermal comfort, and indoor CO_2 levels in comparison to ASHRAE Guideline 36.

4. Methods

This section presents a comprehensive overview of the control strategies implemented in the case study. It details the control sequences governing the HVAC system, which adhere to the G36 standard, and provides an in-depth discussion of the DRL algorithms employed for optimizing system performance.

4.1. ASHRAE guideline 36

ASHRAE Guideline 36–2021 [8] provides standardized sequences of operation for air-handling units (AHUs) serving variable air volume (VAV) terminal systems, optimizing energy efficiency, indoor air quality (IAQ), and control stability. These sequences ensure that HVAC systems dynamically adjust to real-time demand, reducing energy consumption while maintaining thermal comfort. A key feature of G36 is the Trim

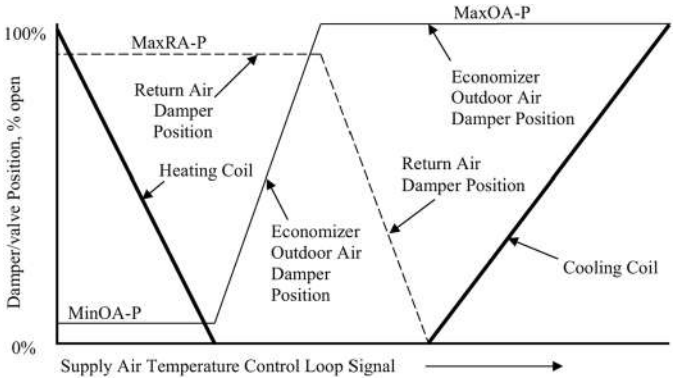


Fig. 9. Mapping of economizer damper positions and coil valve modulation as a function of the Supply Air Temperature (SAT) loop signal [8].

and Respond (T&R) control logic, which enables adaptive, demand-based setpoint adjustments. Unlike static setpoints, this strategy continuously monitors terminal unit performance and dynamically adjusts supply temperature and duct static pressure setpoints according to the system parameters.

The central AHU operation is characterized by the foolowing control variables:

- Supply air temperature (SAT) Reset: the SAT setpoint is dynamically modulated, starting from the minimum cooling value (Min ClgSAT) when the outdoor air temperature (OAT) is at its maximum (OAT Max) and gradually increasing to a defined upper limit (T-max) when the OAT drops to its minimum (OAT Min) Fig. 7. This upper limit, T-max, is further adjusted by T&R logic, which responds to zone reset signals. These are triggered when there is a significant temperature mismatch or prolonged activity in the cooling circuit. If a zone remains above its target temperature for a prolonged period, the system generates additional reset requests that continue until the cooling demand decreases, promoting both temperature control and energy efficiency.
- Duct static pressure reset: the static pressure setpoint is dynamically modulated using the T&R approach, incrementally reducing pressure when all terminal units receive adequate airflow Fig. 8. The system monitors airflow deviations and damper openings, increasing the setpoint when dampers are nearly fully open and airflow is insufficient. The number of reset requests decreases as the deviation lessens, allowing the system to maintain minimum static pressure while meeting zone demand efficiently.
- Economizer control: The economizer damper control system is designed to adjust damper positions to optimise the use of outdoor air for cooling purposes while also ensuring that proper ventilation is maintained. The determination of minimum and maximum damper positions is achieved through the utilisation of airflow and/or pressure sensing mechanisms Fig. 9. The economizer's operation is influenced by various factors, including outdoor temperature, enthalpy, supply fan status, frost protection, and zone conditions. The modulation of dampers is ultimately governed by the SAT setpoint loop.
- Coil valve control: The heating and cooling coil valves are regulated by the SAT control loop, which is governed by a PI controller tracking the SAT setpoint. When the fan is off, the control signal is zero Fig. 9.

The control logic for a VAV reheat terminal unit adjusts damper and valve positions to maintain desired airflow and zone temperature based on the current operating mode.

- Cooling mode: Airflow is modulated between minimum and maximum setpoints to meet cooling demand. The heating coil remains disabled unless the discharge air temperature drops below 10°C, preventing overcooling.
- Heating mode: Airflow is maintained at its minimum setpoint to ensure ventilation, while the heating coil is modulated as needed to meet the zone heating demand. Heating is enabled only when the zone temperature drops below the setpoint.

4.2. DRL control strategies

In this study, two Deep Reinforcement Learning (DRL) strategies based on the Soft Actor-Critic (SAC) framework are employed: *Zone-Aware DRL* and *Zone-Integrated DRL*. A scheme of the DRL control loop is provided in Fig. 10, illustrating the flow of information between the building environment, the DRL agents, and the reward signal. The two strategies differ in their observation space, as summarized in Table 2. *Zone-Aware DRL* monitors each zone temperature at every timestep, while *Zone-Integrated DRL* uses engineered variables representing the total deviation of all zones from the indoor temperature band Eqs. (3) and (4). During occupied periods, the indoor temperature band is defined as a temperature range of ±1 °C around the setpoint temperature (T_{sp}) which is set to 21 °C in the winter season and 26 °C in the summer season. This study examines the impact of the observation granularity on system performance, assessing whether an aggregated approach can achieve comparable efficiency to a zone-specific strategy.

Despite these differences in the observation space, both approaches control the following actuators with a control timestep of 15 minutes.

- Fan speed (fan_{speed}): Adjusting the speed of the supply fan to regulate airflow.
- Outdoor air damper position (dam_{out}): Modulating the outdoor air intake damper to control fresh air intake.
- Return air damper position (dam_{ret}): Adjusting the return air damper to balance recirculated and fresh air.
- Coil valve opening ($coil_{\text{valve}}$): Regulating the heating or cooling coil valve to manage the supply air temperature.

Instead, the Variable Air Volume (VAV) box controls follow the G36 guideline as described in Section 4.1. The reward function differs between the cooling and heating seasons, as outlined in Eqs. (1) and (2). Specifically, the term related to energy consumption changes during the heating season, where the total energy cost must account for the post-heating demand across all five VAV boxes. This additional term was introduced to prevent biased and unrealistic control strategies, ensuring that the agent does not minimize AHU energy use by shifting heating loads to the VAV boxes. Representing the post-heating demand through a single aggregated term also keeps the reward formulation fair and as simple as possible, given its multi-objective nature. The weights of the reward terms (λ_{fan}, λ_{HP}, λ_{up}, λ_{down}, λ_{CO_2}, and λ_{VAV}) were selected to balance the relative influence of comfort and energy objectives. They were initially set according to the order of magnitude of the corresponding variables and subsequently refined through iterative trial-and-error testing. The tuning process has stopped once the DRL controllers achieved a reasonable trade-off, yielding about a 17 % reduction in energy consumption compared to the baseline, while maintaining acceptable indoor temperature violations and CO_2 levels.

$$r_{\text{cooling}} = -\begin{cases} \lambda_{HP}E_{HP}^2 + \lambda_{fan}E_{fan}^2 + \lambda_{up}ZAT_{up}^2 + \lambda_{\text{down}}ZAT_{down}^2 \\ \quad +\lambda_{\text{CO}_2}\max\left(0, \text{CO}_2 - \text{CO}_{2_{\max}}\right)^2, \text{if occupied} \\ \lambda_{HP}E_{HP}^2 + \lambda_{\text{CO}_2}\max\left(0, \text{CO}_2 - \text{CO}_{2_{\max}}\right)^2, \text{if unoccupied} \end{cases} \quad (1)$$

$$r_{\text{heating}} = r_{\text{cooling}} - \lambda_{VAV}\sum_{vav=1}^{N_{\text{zones}}} E_{VAV} \quad (2)$$

where:

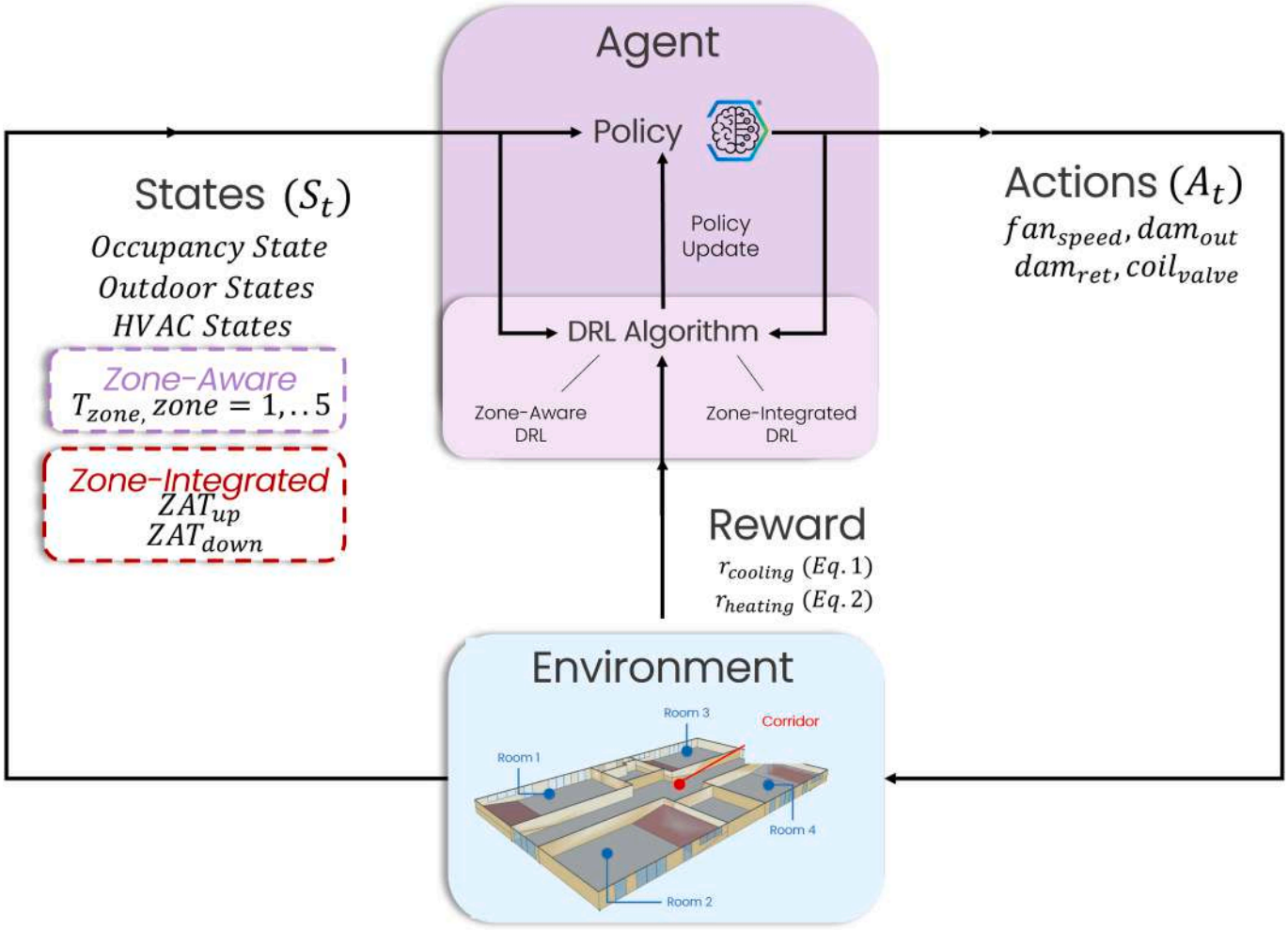


Fig. 10. Structural visualization of the DRL control framework. The agent interacts with the building environment by observing occupancy, outdoor, and HVAC states together with zone information, and selecting AHU-level actions (fan speed, outdoor/return damper positions, coil valve opening). The environment responds with the next state and a scalar reward reflecting energy use, indoor temperature violations, and CO_2 levels, enabling the agent to iteratively improve its control policy. The observation space differs between Zone-Aware and Zone-Integrated strategies as detailed in Table 2.

Table 2
Observation space comparison of zone-aware and zone-integrated DRL agents.

Variable	Description	Zone-Aware	Zone-Integrated
T_{out}	Outdoor air temperature	✓	✓
T_{ret}	Return air temperature	✓	✓
T_{mix}	Mixed air temperature	✓	✓
T_{sup}	Supply air temperature	✓	✓
Solar gain	Solar radiation gain	✓	✓
T_{zone}, zone = 1.5	Zone temperatures	✓	×
ZAT_{up}	Total zones' violation above upper bound	×	✓
ZAT_{down}	Total zones' violation below lower bound	×	✓
V_{TOT}	Total ventilation flow rate	✓	✓
V_{out}	Outdoor airflow rate	✓	✓
CO_2	Indoor CO_2 concentration	✓	✓
E_{fan}	Electric energy used by the fan	✓	✓
E_{HP}	Electric energy used by heat pump/chiller	✓	✓
COP_{HP}	Coefficient of performance of heat pump/chiller	✓	✓
$t_{\text{start\_occupied}}$	Time until occupancy starts	✓	✓
$t_{\text{end\_occupied}}$	Time until occupancy ends	✓	✓
$T_{\text{out\_1h}}$	Forecasted outdoor temperature (1h)	✓	✓
$T_{\text{out\_3h}}$	Forecasted outdoor temperature (3h)	✓	✓
$T_{\text{out\_6h}}$	Forecasted outdoor temperature (6h)	✓	✓

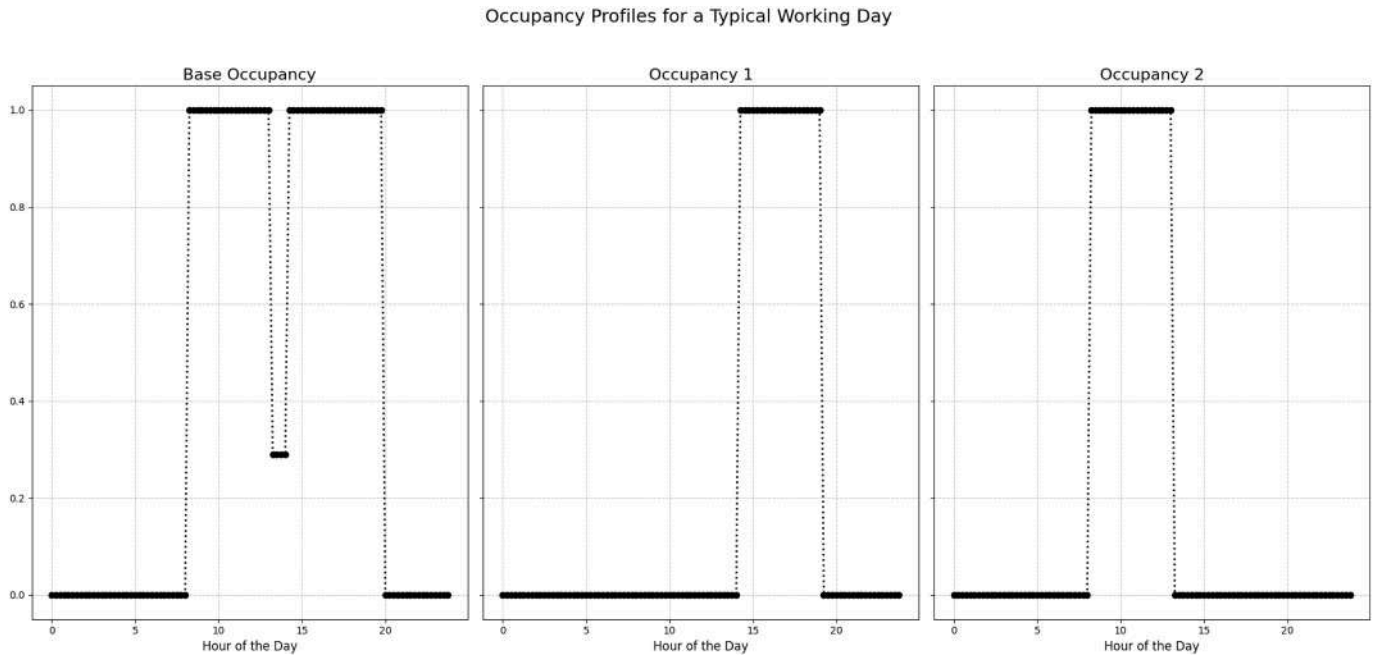


Fig. 11. Occupancy profiles representing percentage of rooms occupation throughout a typical working day (Monday to Friday).

- ZAT_{up} represent the zone air temperature violation from the upper bound of the band ($T_{sp}+1$) and is given by the relation

$$ZAT_{up} = \sum_{zone=1}^{N_{\text{zones}}} \max(0, T_{zone} - (T_{sp}+1)) \tag{3}$$

- ZAT_{down} represent the zone air temperature violation from the lower bound of the band ($T_{sp}-1$) and is given by the relation

$$ZAT_{down} = \sum_{zone=1}^{N_{\text{zones}}} \max(0, (T_{sp}-1) - T_{zone}) \tag{4}$$

- E_{HP} is the energy consumption of the heat pump/chiller.
- E_{VAV} accounts for the additional heating demand in all five VAV boxes.
- E_{fan} is the energy consumption of the fan.
- $CO_{2_{\max}}$ is the maximum allowed value of CO_2 set to 1000 ppm.

5. Results

This section presents a comprehensive performance evaluation of the Zone-Aware DRL and Zone-Integrated DRL controllers, compared to the ASHRAE G36 baseline, under varying boundary conditions such as weather and occupancy profiles. The analysis focuses on the following distinct occupancy patterns (see also Fig. 11), evaluated during both winter and summer periods to assess performance under typical conditions and to test the controllers' ability to adapt to previously unseen scenarios.

- **Occupancy 0 (Base case)**: this profile represents the nominal training condition, characterized by continuous occupancy from 08:00 to 19:45, Monday through Friday. This profile reflects standard campus operating hours and serves as the baseline for evaluating model performance under seen conditions.
- **Occupancy 1**: this profile simulates partial afternoon occupancy from 14:00 to 19:00, Monday through Friday. This profile is not included in the training dataset and it is used to evaluate generalization to unseen temporal patterns.
- **Occupancy 2**: this profile represents partial morning occupancy from 08:00 to 13:00, Monday through Friday. Like Occupancy 1, this scenario is not seen during training and provides additional insight into controller adaptability.

For both winter and summer scenarios, the training was carried out exclusively in the Occupancy 0 profile, representing the full-day occupancy condition (08:00-19:45). Winter training used data from January 1, to January 31, with an evaluation that spanned January 1, to February 28. Similarly, summer training was conducted from 1 July to 31 July, with evaluation covering the period from 20 June to 20 August.

The analysis of Occupancy 1 and 2 focuses on the controllers' ability to adapt their AHU control strategies–particularly in terms of scheduling and operational efficiency–in response to previously unseen occupancy dynamics. This provides critical insight into the robustness, flexibility, and real-world applicability of the learned control policies under realistic, time-varying building usage.

Performance is assessed using three key metrics aligned with the reward functions design (Eqs. (1) and (2)): (i) total energy consumption, (ii) indoor temperature violations at the zone level, and (iii) average indoor CO_2 concentration. These indicators jointly reflect the controllers' ability to balance energy efficiency, occupant comfort, and indoor air quality.

5.1. Performance under nominal occupancy profile (occupancy 0)

During winter, DRL-based controllers significantly improve energy performance by shifting the heating strategy away from inefficient terminal VAV reheating and toward centralized heating provided by the AHU heat pump. Compared to G36, both the Zone-Aware and Zone-Integrated DRLs reduce VAV energy consumption by approximately 24 % by decreasing the reliance on local terminal reheating and instead meeting a larger share of the heating demand through the central AHU system. This redistribution is reflected in the increased heat pump usage, with DRLs consuming up to 40 % more central heating energy. However, this increase is offset by overall efficiency gains, producing net reductions in total winter energy use - 16 % for the Zone-Aware controller and 14.4 % for Zone-Integrated Table 3.

Further insight into heating behavior is revealed through thermal energy and heat pump-hour analysis Fig. 12(a). Although DRLs achieve lower total thermal energy demand than G36, their heat pumps are

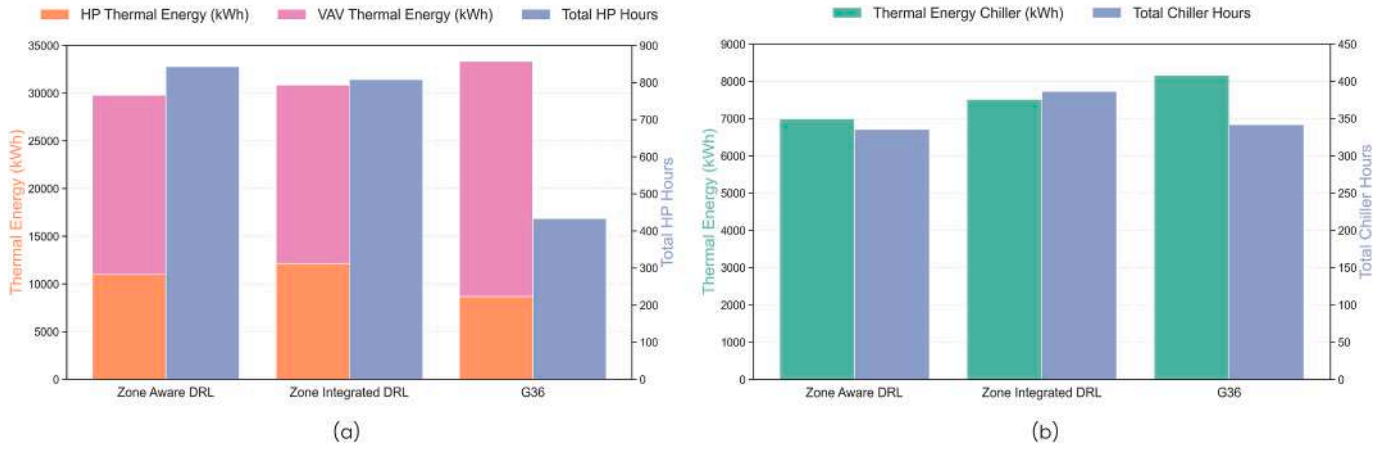


Fig. 12. Total thermal energy delivered and heat pump operating hours during (a) the winter season (January 1st to February 28th) and (b) the summer season (June 20th to August 20th).

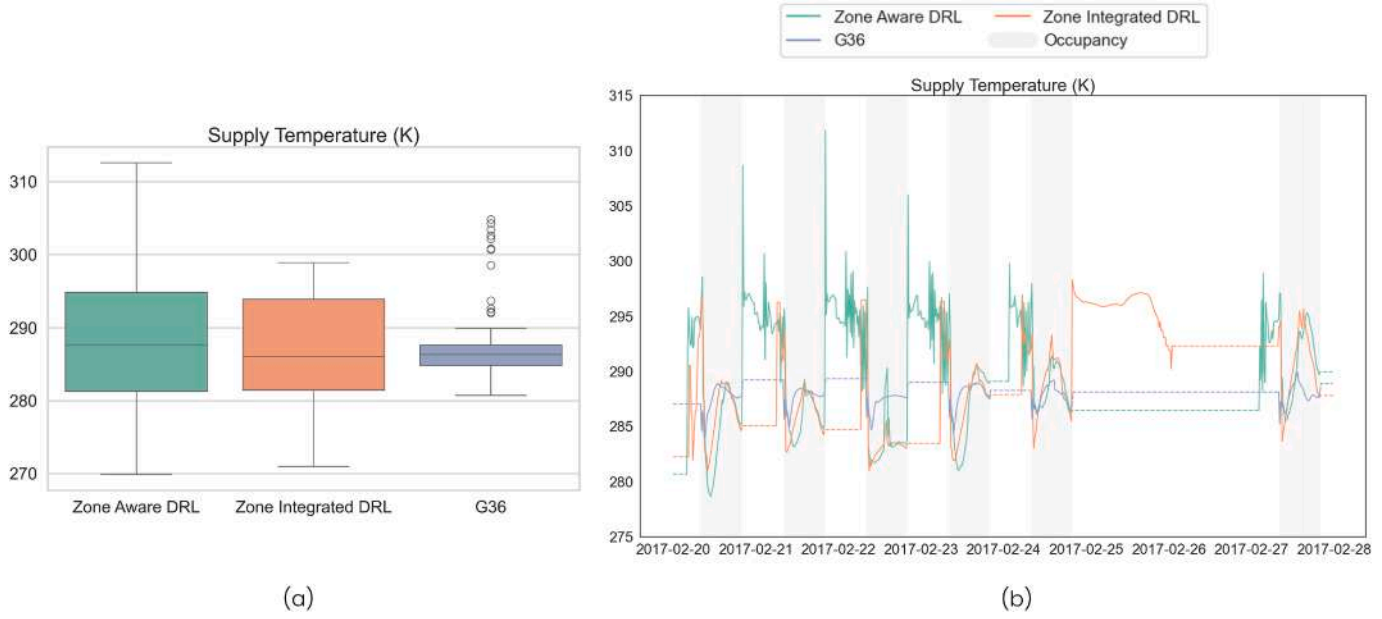


Fig. 13. Supply temperature (T_{sup}) by control strategy: (a) boxplot distribution during the winter season (January 1st to February 28th), and (b) temporal pattern over a representative winter week (dashed lines indicate periods of system inactivity).

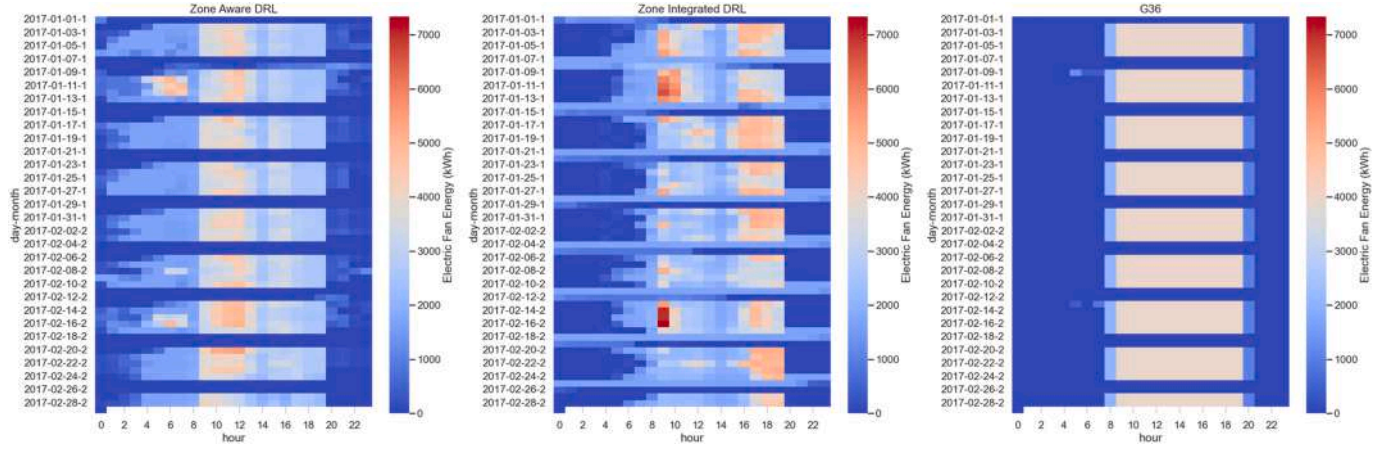


Fig. 14. Hourly fan electricity consumption and activation patterns for different control strategies during the winter season.

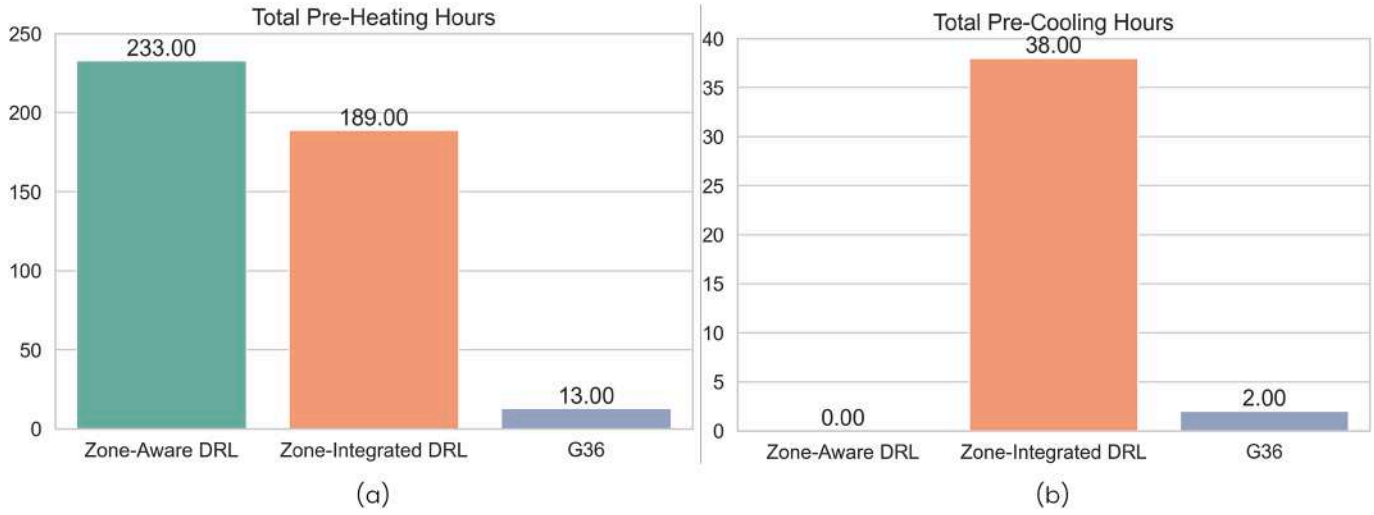


Fig. 15. Total hours of pre-conditioning under each control strategy: (a) pre-heating during the winter season (January 1st to February 28th), and (b) pre-cooling during the summer season (June 20th to August 20th).

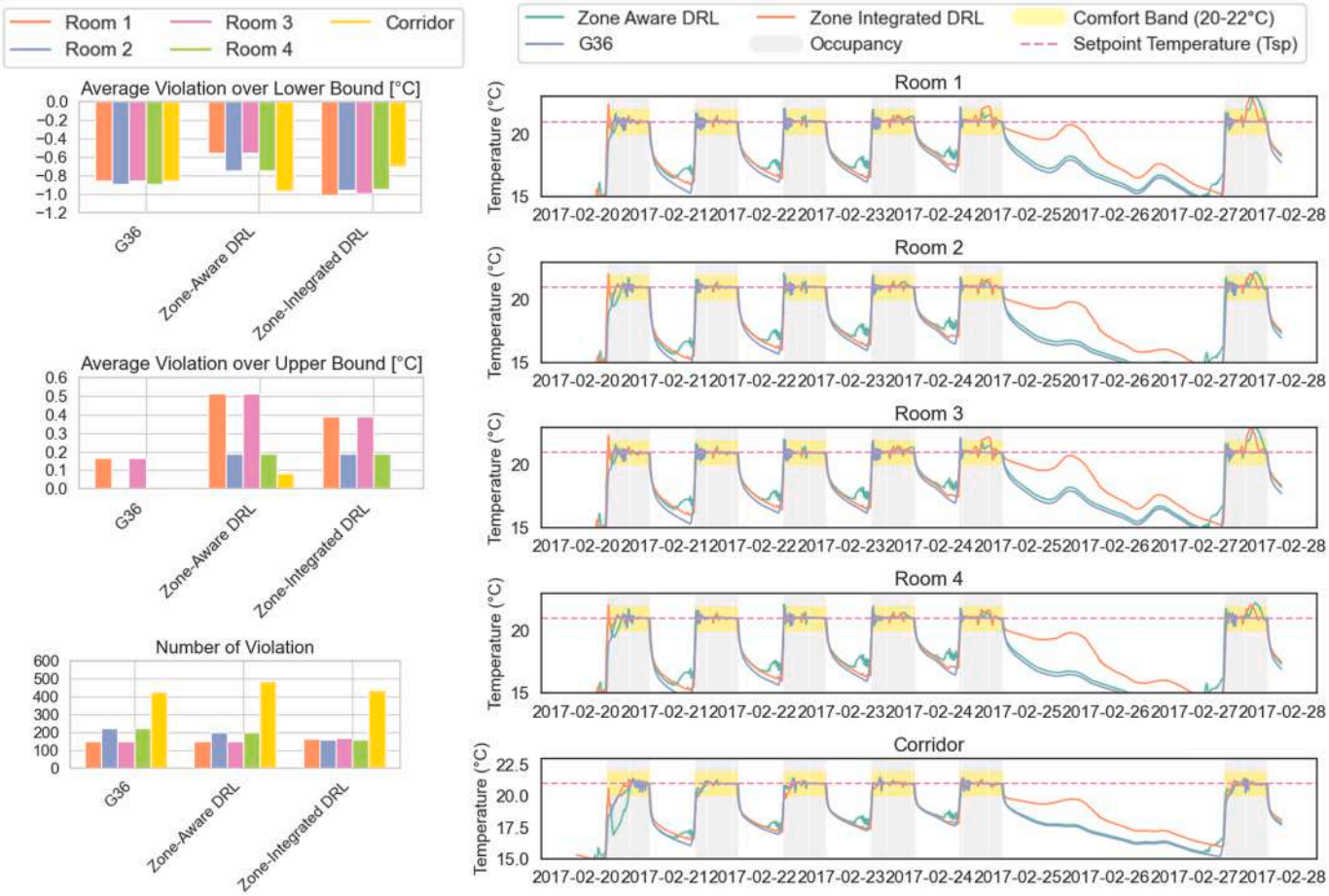


Fig. 16. Indoor temperature band violations during the winter period (January 1st to February 28th) for each zone and control strategy (left), and temperature patterns over a representative winter week for each zone and strategy (right).

active for nearly twice as long. The Zone-Aware and Zone-Integrated controllers reach around 860 and 820 heat pump operation hours, respectively, compared to just 440 hours for G36. This suggests that DRLs prefer longer-duration, low-intensity heating–a strategy enabled by more precise coordination between central delivery and airflow. This behavior is consistent with supply air temperature patterns Fig. 13: while G36 maintains a nearly fixed T_{sup}, DRLs exhibit broader and more dynamic ranges. Notably, the Zone-Aware DRL demonstrates more anticipatory behavior, often ramping up supply temperatures earlier to precondition spaces ahead of occupancy, while the Zone-Integrated DRL exhibits a slightly more reactive pattern, in accordance with the total hours of pre-heating of the different strategies Fig. 15(a). These

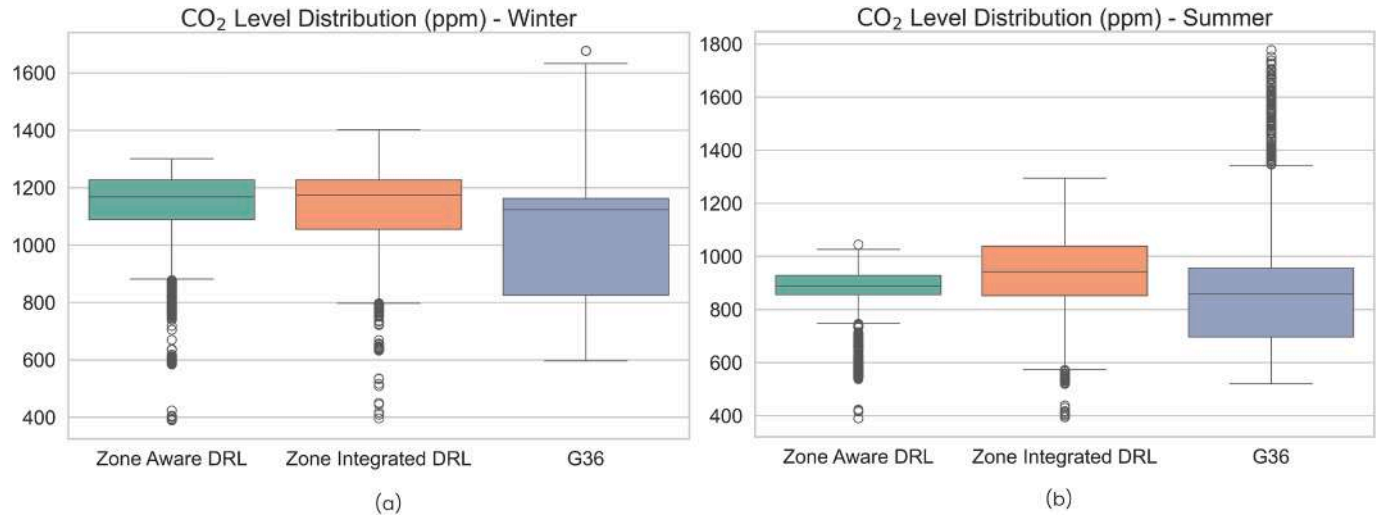


Fig. 17. Boxplot distribution of indoor CO_2 concentrations during (a) the winter period (January 1st to February 28th) and (b) the summer period (June 20th to August 20th).

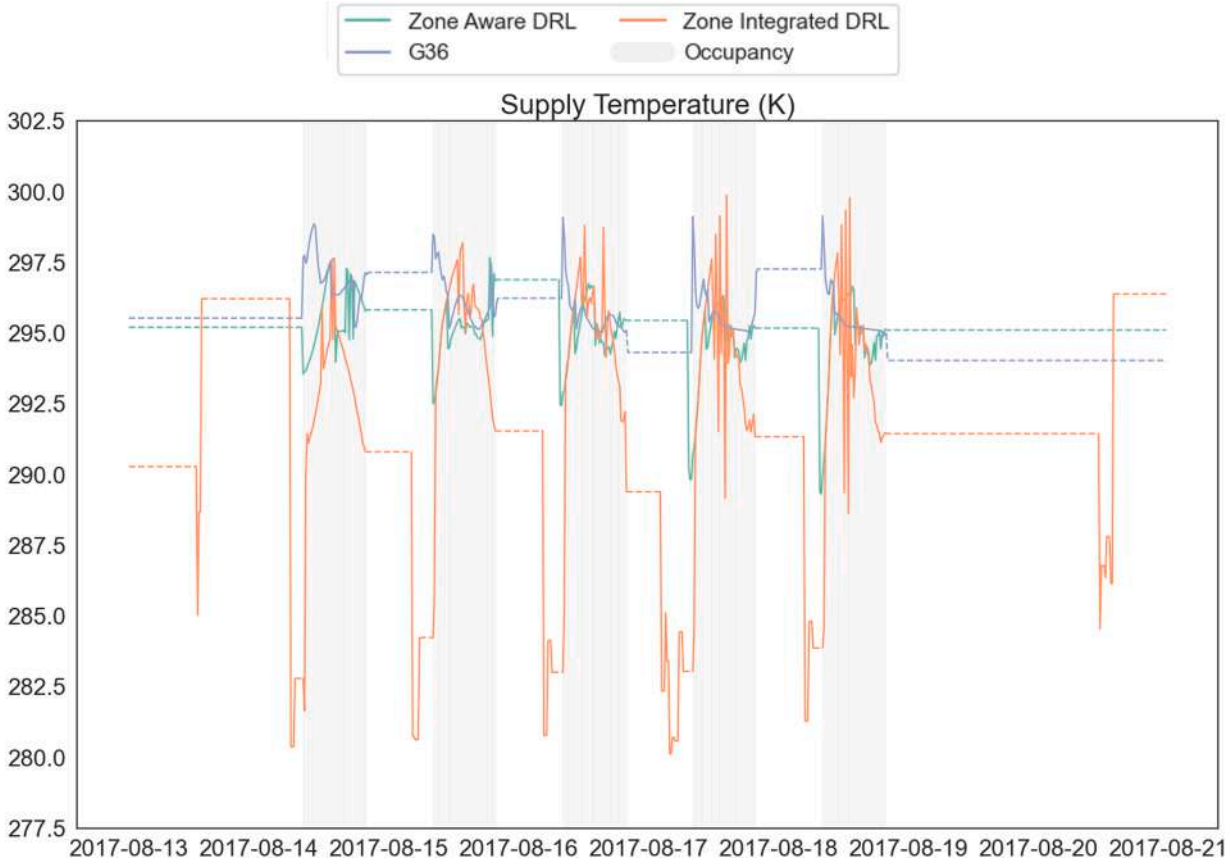


Fig. 18. Temporal pattern of supply temperature (T_{sup}) over a representative summer week for each control strategy (dashed lines indicate periods of system inactivity).

approaches allow both DRLs to leverage the heat pump's higher coefficient of performance (COP) instead of relying on less efficient VAV reheating.

Fan operation is also adapted to support this shift: DRLs operate fans for longer durations but at lower speeds, exploiting the cubic fan power curve to maintain efficiency. This distributed and modulated airflow is visible in the fan energy heatmaps, where DRLs contrast sharply with G36's rigid, time-constrained daytime operation Fig. 14. Despite these operational shifts, both DRL strategies maintain thermal comfort within acceptable ranges Fig. 16. Regarding indoor CO_2 concentrations Fig. 17(a), the Zone-Aware DRL exhibits the narrowest distribution with minimal high-end outliers, indicating more consistent and responsive ventilation aligned with occupancy. The Zone-Integrated DRL also maintains acceptable air quality, though with slightly greater variability across the distribution. In contrast, the baseline G36, although having a slightly lower median, shows a wider spread and several high outliers,

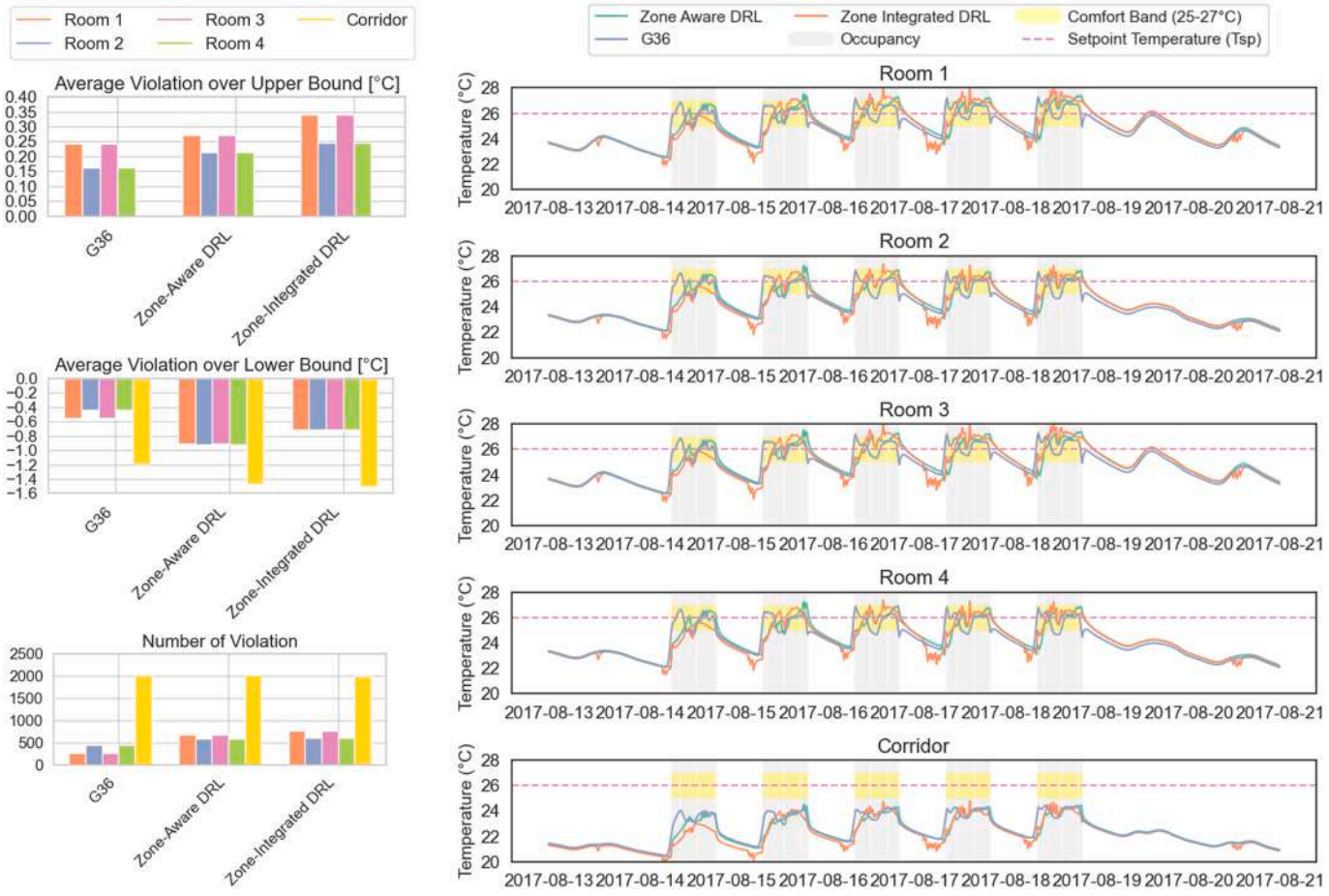


Fig. 19. Indoor temperature band violations during the summer period (June 20th to August 20th) for each zone and control strategy (left), and temperature patterns over a representative summer week for each zone and strategy (right).

with concentrations occasionally exceeding 1600 ppm. These peaks occur primarily during early occupancy periods, likely due to delayed ventilation activation caused by the fixed control schedule of G36. Considering a CO_2 generation rate of 20 L/h per person, and referencing the thresholds outlined in [45], all control strategies fall within Category III air quality.

In the summer period, DRLs continue to outperform G36 by reducing fan and cooling energy through more adaptive, demand-responsive operation. Both DRLs achieve approximately 17 % total energy savings Table 3, with the Zone-Aware controller again demonstrating the lowest cooling load with the fewest heat pump activation hours Fig. 12(b). G36 delivers 8,158 kWh of thermal energy through the cooling coil, while the DRLs reduce this to 7,512 kWh (Zone-Integrated) and 6,997 kWh (Zone-Aware). These reductions, when considered alongside chiller runtime, reveal deeper efficiency gains: the Zone-Integrated DRL operates the chiller for the longest duration (~395 hours) but still consumes less energy than G36, indicating lower cooling intensity per hour and suggesting partial load coverage through early activation and free cooling. This is further supported by the pre-conditioning analysis Fig. 15(b), which shows the Zone-Integrated controller performing over 38 hours of pre-cooling–substantially more than both G36 (1.6 hours) and the Zone-Aware controller (0 hours). Conversely, the Zone-Aware DRL achieves similar savings with even fewer coil hours (~335), highlighting a strategy that combines passive cooling and load tracking.

This adaptive behavior is further reflected in the supply air temperature (T_{sup}) profiles Fig. 18. While G36 maintains T_{sup} within a narrow, mostly static range throughout the day, both DRL strategies adapt T_{sup} with much greater flexibility. The Zone-Aware DRL modulates T_{sup} smoothly during occupancy periods with minor frequent adjustments that suggest demand tracking. On the other hand, the Zone-Integrated DRL exhibits sharp drops in T_{sup}–particularly in early morning hours–even before occupancy begins, a behavior indicative of proactive pre-cooling. These early intense cooling, followed by higher and more variable values of T_{sup} during peak hours, suggest a strategy that cools the building in advance to reduce the need for cooling later in the day.

Although indoor temperature violations slightly increase under DRL control, they remain localized in secondary spaces (e.g., corridors) and within acceptable thresholds Fig. 19. Meanwhile, CO_2 concentrations are consistently maintained within target limits, ensuring adequate indoor air quality Fig. 17(b).

Table 3
Electrical energy consumption breakdown (kWh) for each controller in winter and summer.

Control Strategy	Winter				Summer		
	HP	Fan	VAVs	Tot.	HP	Fan	Tot.
G36	3521	1963	24,665	30,149	2016	2031	4047
Zone-Aware DRL	4505	2009	18,834	25,348	1701	1640	3341
Zone-Integrated DRL	4929	2143	18,741	25,813	1761	1586	3347

5.2. *Unseen occupancy profiles*

5.2.1. *Heating season*

Unlike the DRL-based strategies, which were trained under the Base Occupancy profile and maintained good performance under both

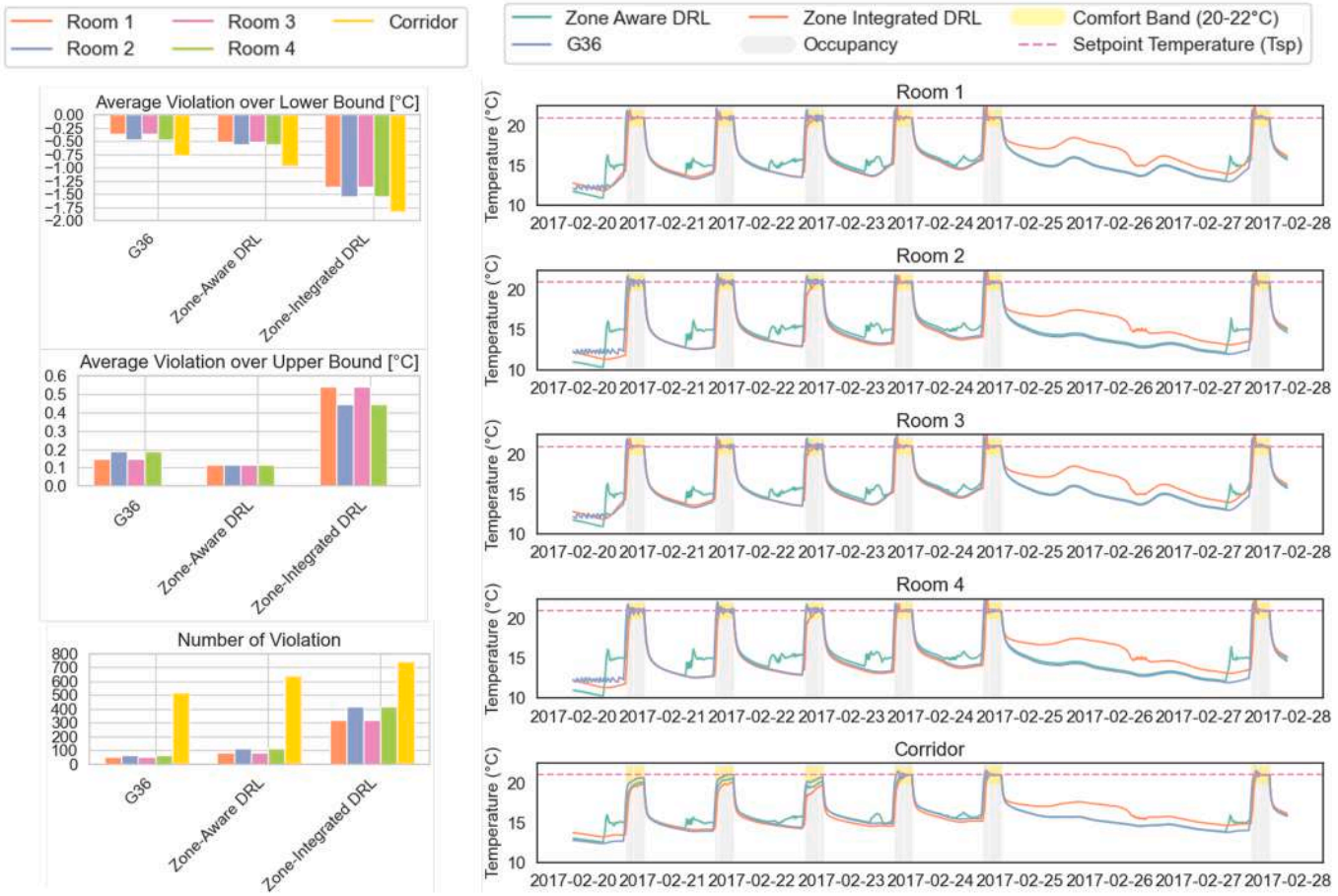


Fig. 20. Indoor temperature band violations during the winter period (January 1st to February 28th) for each zone and control strategy (left), and temperature profiles over a representative winter week (right), under Occupancy Profile 1 (14:00-19:00).

modified winter schedules (Occupancy 1 and Occupancy 2) without the need for further fine-tuning, the G36 baseline required manual recalibration to accommodate each new occupancy configuration. This highlights the ability of DRL methods to adapt autonomously to changing operating conditions.

Focusing first on the Zone-Aware DRL, its anticipatory heating behavior plays a central role in both schedules. Under Occupancy 1 (14:00-19:00), this results in a modest increase in energy consumption (+3.8 %) compared to G36, driven by early activation to precondition spaces before occupancy begins. It is important to note, however, that all controllers–including G36–are subject to system-level requirements that allow heating during unoccupied periods if zone temperatures fall below 15 °C. This condition contributes to early heating, particularly in the afternoon schedule where extended unoccupied periods coincide with lower temperatures. However, this behavior aligns well with thermal demand under Occupancy 2 (08:00-13:00), where the controller achieves the lowest total energy use among all strategies while maintaining thermal comfort equivalent to G36 Table 4.

The Zone-Integrated DRL adopts a more reactive control pattern. Under Occupancy 1, this allows a slight improvement with respect to G36 in energy consumption (-0.4 %) by avoiding early heating, though at the cost of increased temperature violations during transitional periods. Under Occupancy 2,while not as efficient as Zone-Aware, it matches G36 in energy use and adapts to the new schedule without manual adjustment–demonstrating robustness and generalizability Table 4.

In terms of indoor temperature control, both DRL strategies maintain acceptable conditions across thermal zones under both schedules Figs. 20 and 21. Zone-Aware achieves violation levels comparable to G36 under Occupancy 1 and outperforms it under Occupancy 2, while Zone-Integrated shows slightly more frequent deviations due to its delayed response behavior.

Table 4
Electrical energy consumption breakdown (kWh) for each control strategy and occupancy profile during the winter period (January 1st to February 28th).

Control Strategy	Occupancy 1				Occupancy 2			
	HP	Fan	VAVs	Tot.	HP	Fan	VAVs	Tot.
G36	2072	1002	11,741	14,815	2543	1023	13,319	16885
Zone-Aware DRL	2705	1035	11,652	15,392	2255	1112	12,871	16,238
Zone-Integrated DRL	2660	1207	10,889	14,756	3170	1348	12,484	17,002

Regarding indoor air quality, all strategies keep CO_2 concentrations within acceptable bounds Fig. 22. However, both DRL controllers exhibit tighter distributions and lower peak values than G36, particularly under Occupancy 2, suggesting more effective and demand-driven ventilation control even as energy consumption is reduced.

5.2.2. *Cooling season*

In the summer scenario, both DRL-based strategies demonstrate adaptability and good performance compared to the G36 baseline, which, as done in the winter scenario, has been recalibrated to adapt to the new occupancy schedules.

The Zone-Aware DRL demonstrates consistently strong performance across both summer occupancy scenarios Table 5. Under Occupancy 1 (14:00-19:00), where cooling demand aligns with peak afternoon

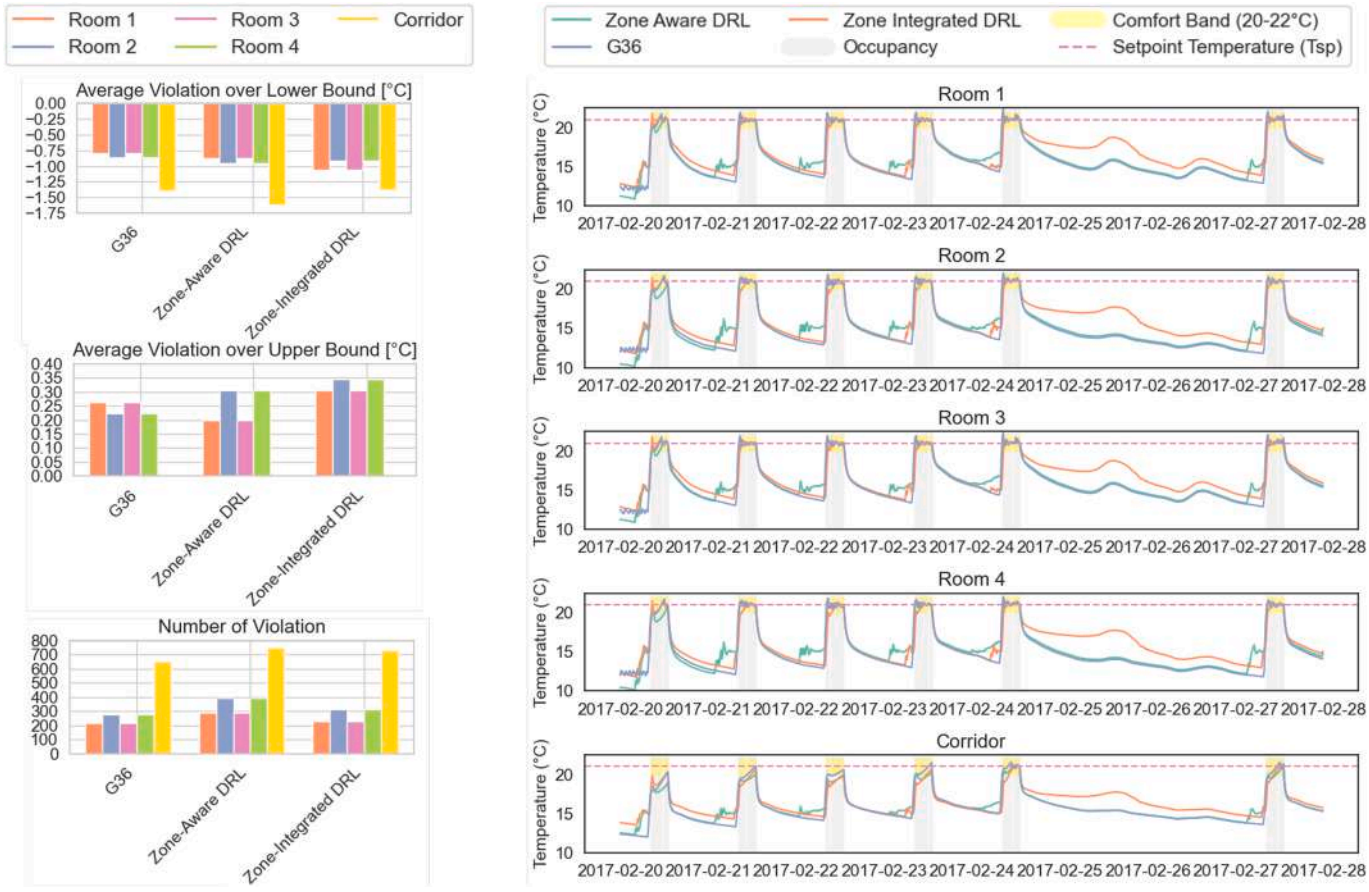


Fig. 21. Indoor temperature band violations during the winter period (January 1st to February 28th) for each zone and control strategy (left), and temperature profiles over a representative winter week (right), under Occupancy Profile 2 (8:00-13:00).

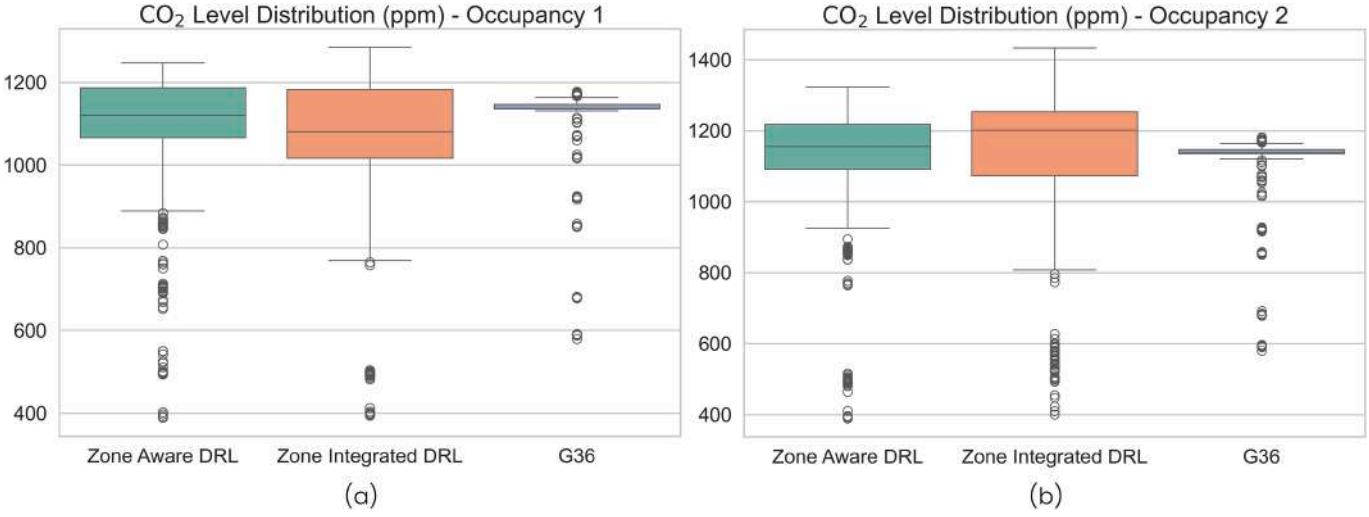


Fig. 22. Boxplot distribution of indoor CO_2 concentrations during the winter period (January 1st to February 28th) under (a) occupancy profile 1 (14:00-19:00) and (b) occupancy profile 2 (8:00-13:00).

outdoor temperatures, it emerges as the most energy-efficient strategy, with around 19 % reduction compared to the G36 baseline. This efficiency is driven by its context-aware control policy that dynamically adjusts system operation based on real-time thermal demand. Zone-Aware minimizes unnecessary activations and finely tunes the supply air temperature (T_{sup}) throughout the day–resulting in a strategy that is both responsive and resource-conscious, as also observed in the fan energy heatmaps Fig. 23. Under Occupancy 2 (08:00-13:00), which takes place during the cooler morning hours, the Zone-Aware DRL reduces energy by approximately 16 % improvement over G36. The controller's ability to dynamically modulate T_{sup} aligns well with lower morning loads and milder ambient temperatures, allowing for comfort to be maintained.

In contrast, the Zone-Integrated DRL, which incorporates pre-cooling as a core behavioral feature, shows more variable performance Table 5. In Occupancy 1, it performs poorly due to its rigid pre-conditioning schedule. The controller activates the system significantly ahead of

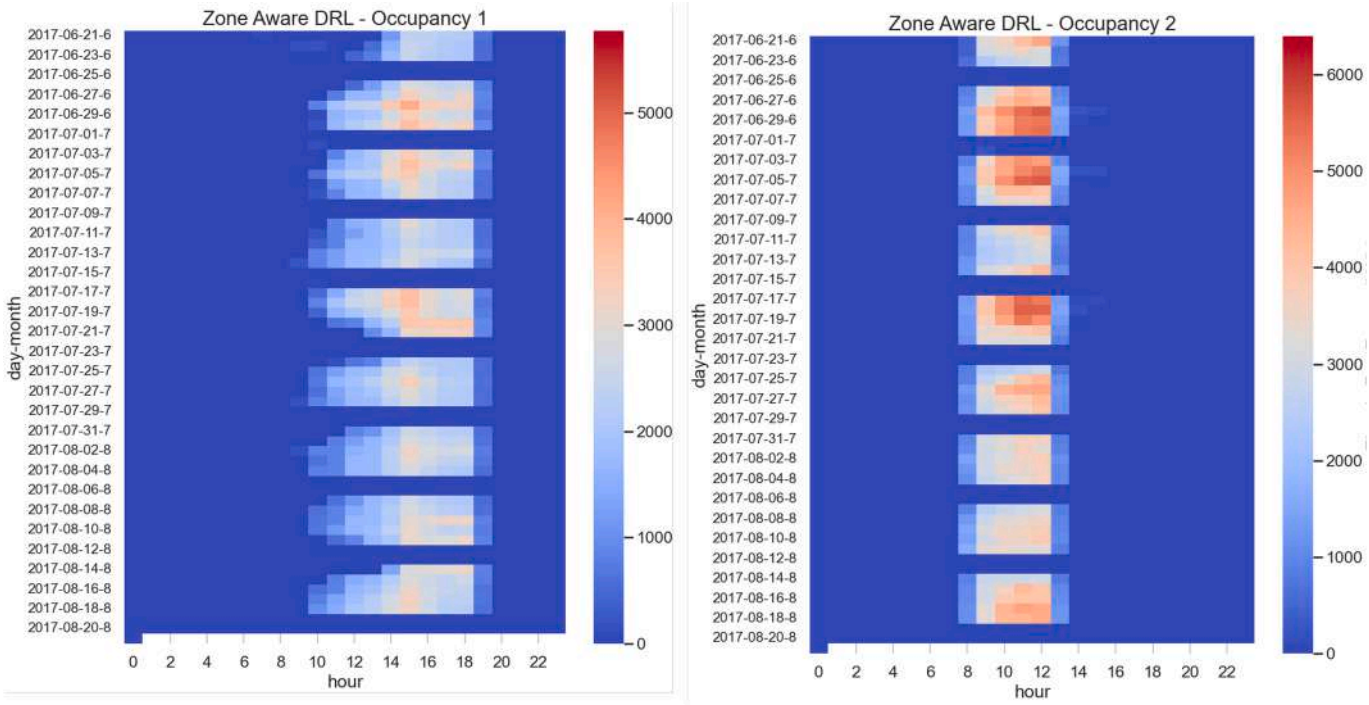


Fig. 23. Fan energy usage during the summer period (June 20th to August 20th) for Occupancy Profile 1 (left) and Occupancy Profile 2 (right).

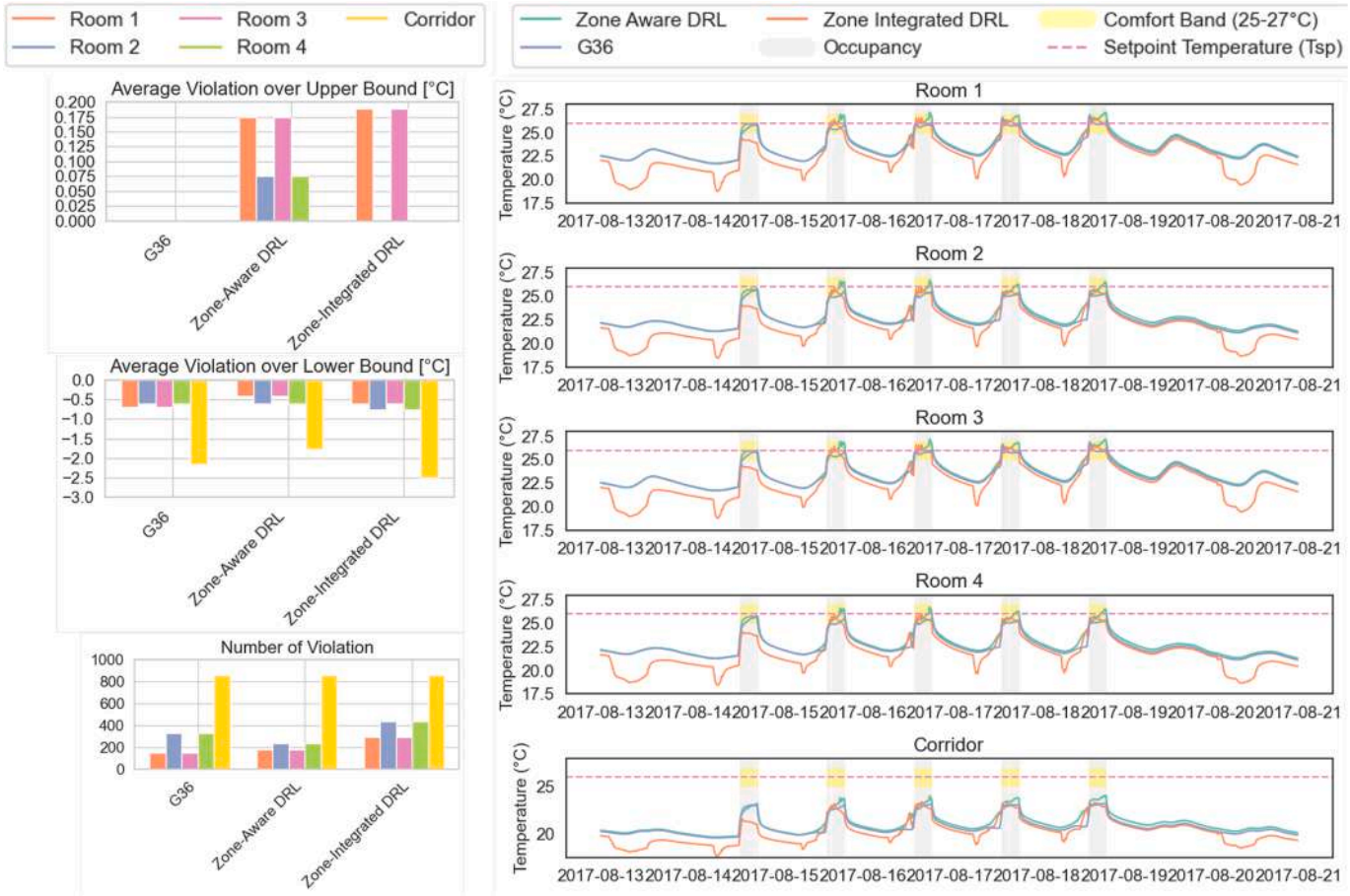


Fig. 24. Indoor temperature band violations during the summer period (June 20th to August 20th) for each zone and control strategy (left), and temperature profiles over a representative summer week (right), under Occupancy Profile 1 (14:00-19:00).

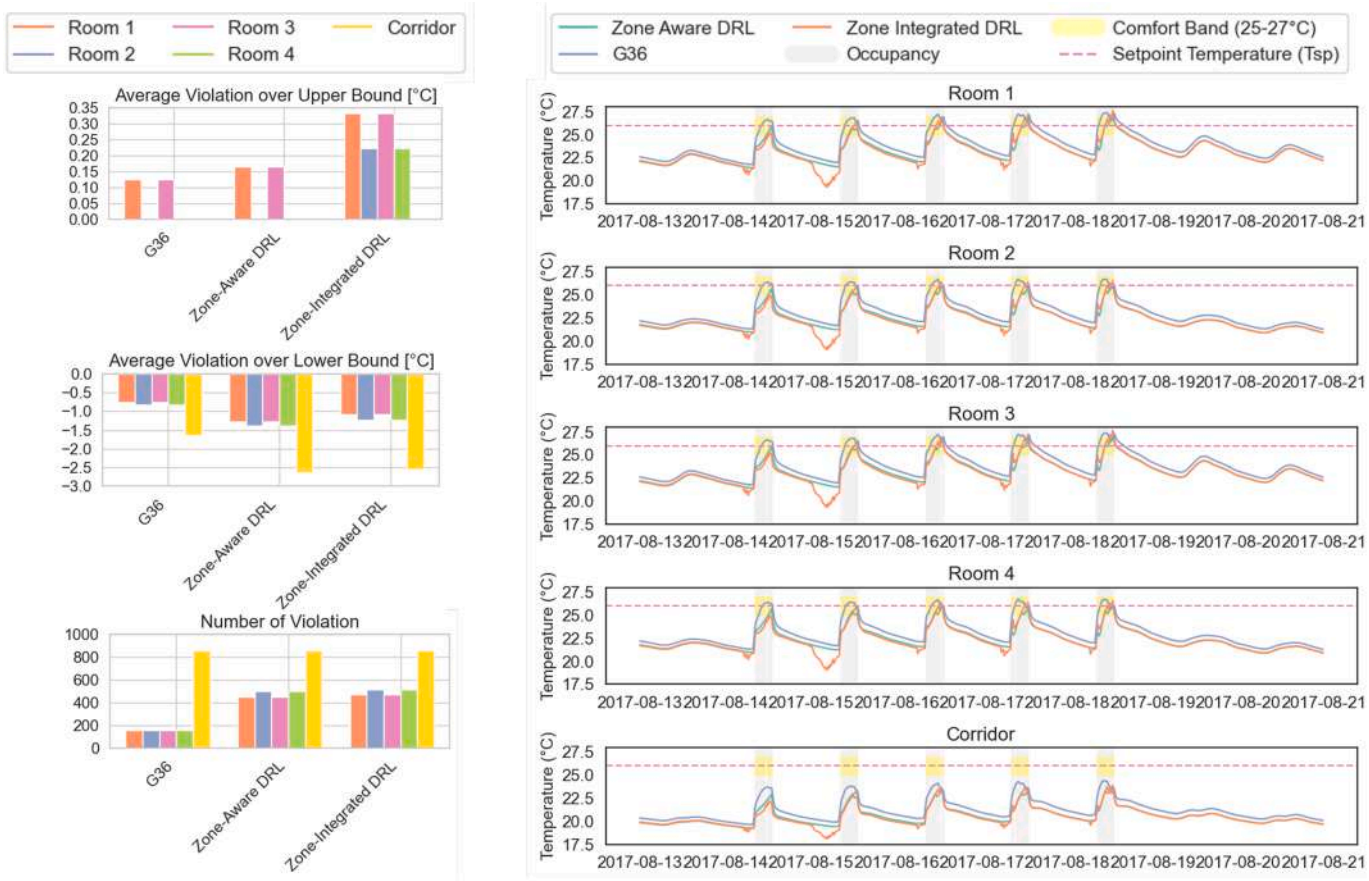


Fig. 25. Indoor temperature band violations during the summer period (June 20th to August 20th) for each zone and control strategy (left), and temperature profiles over a representative summer week (right), under Occupancy Profile 2 (8:00-13:00).

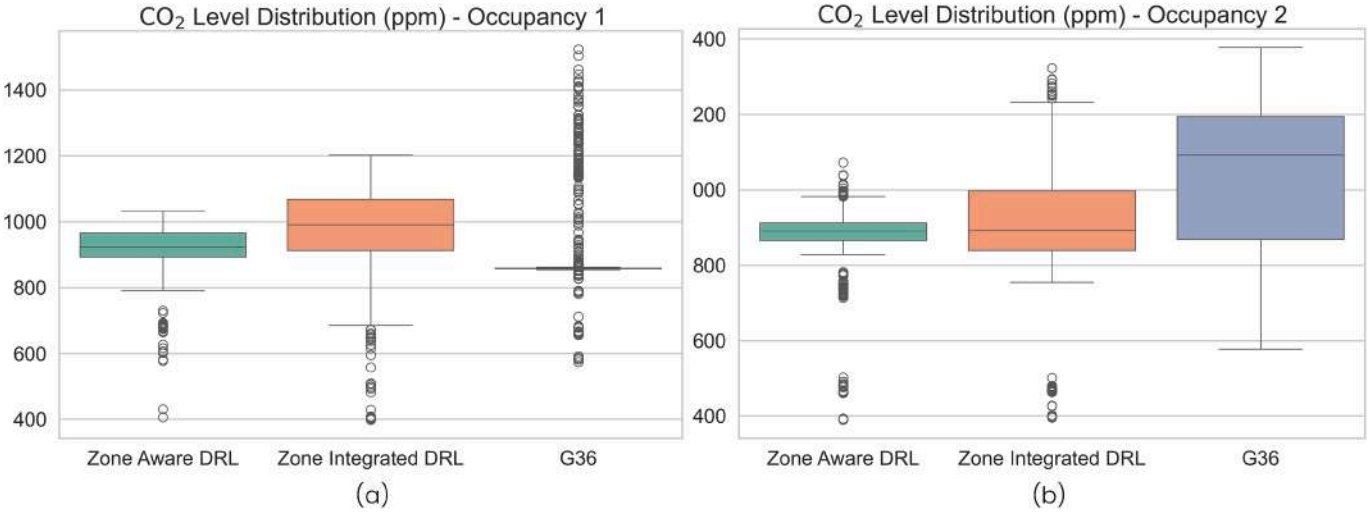


Fig. 26. Boxplot distribution of indoor CO_2 concentrations during the summer period (June 20th to August 20th) under (a) Occupancy Profile 1 (14:00-19:00) and (b) Occupancy Profile 2 (8:00-13:00).

occupancy, accumulating over 74 hours of pre-cooling time. This early operation is ineffective under high late morning/early afternoon outdoor temperatures and leads to excessive use of the chiller (1254 kWh), ultimately yielding the highest total energy consumption among the strategies at 2186 kWh. However, under Occupancy 2, the same pre-cooling behavior becomes more suitable, given the cooler morning hours. In this case, the controller reduces energy use to 1024 kWh–approximately 10 % less than G36–though it still underperforms relative to the more flexible and context-aware Zone-Aware DRL.

Despite differences in control behavior, both DRL strategies maintain acceptable levels of thermal comfort. Under Occupancy 1, the Zone-Aware DRL delivers comfort performance comparable to G36, with only minor increases in upper temperature violations. The Zone-Integrated DRL exhibits more frequent and severe comfort violations, particularly during late-day transitions when pre-cooling is less effective Fig. 24.

Table 5
Electrical energy consumption breakdown (kWh) for each controller and occupancy profiles in summer.

Control Strategy	Occupancy 1			Occupancy 2		
	HP	Fan	Tot.	HP	Fan	Tot.
G36	1168	823	1991	328	815	1143
Zone-Aware DRL	831	784	1615	223	734	957
Zone-Integrated DRL	1254	932	2186	241	783	1024

Under Occupancy 2, comfort metrics improve across all strategies. Nevertheless, Zone-Aware remains the most consistent performer, minimizing both the frequency and magnitude of violations. Zone-Integrated, while improved, still displays slightly greater variability in maintaining setpoint conditions across zones Fig. 25.

Indoor air quality, as assessed through CO_2 concentration, further illustrates the strengths of learning-based control Fig. 26. Zone-Aware achieves the lowest median CO_2 levels and the narrowest distribution across both occupancy scenarios, indicating precise and timely ventilation aligned with real occupancy patterns. The Zone-Integrated DRL also maintains CO_2 within acceptable thresholds, although with broader variability–especially under Occupancy 1–reflecting less consistent ventilation timing. In contrast, the G36 baseline shows the highest CO_2 variability and more frequent excursions above recommended limits, a consequence of its rigid, schedule-based ventilation strategy that lacks responsiveness to dynamic indoor conditions.

6. Discussion

The results show the effectiveness of DRL-based controllers in optimizing HVAC performance relative to the ASHRAE G36 baseline. When evaluated under the same occupancy schedule used for training, both DRL strategies outperform the G36 guideline across winter and summer seasons. The Zone-Aware controller consistently achieves the highest energy savings while maintaining thermal comfort and indoor CO_2 level. However, the Zone-Integrated DRL also performs competitively, with only marginally lower efficiency in most metrics. This suggests that a controller which leverages global system-level variables, rather than zone-specific feedback, can still yield effective performance.

When tested under unseen occupancy profiles the G36 baseline required manual retuning to adapt to the modified schedules, while both DRL controllers adapted autonomously–without any policy adjustment. This underscores the flexibility of learned control strategies.

In the heating season, the generalization capabilities of both DRLs remain within the performance range of G36. However, the margin for optimization is inherently narrower in winter due to the dominant role of the VAV terminals, which tightly track zone-level temperature setpoints regardless of occupancy variability. As a result, the DRLs' flexibility is somewhat constrained, and their improvements over the baseline are more modest. In contrast, the cooling season reveals a more robust and consistent generalization of the DRL policies–particularly for the Zone-Aware controller. Under both unseen occupancy profiles, it adapts effectively by modulating supply air temperature and ventilation to match dynamic load conditions. Its context-aware operation enables significant energy savings without compromising occupant comfort or air quality. The Zone-Integrated DRL, though somewhat less responsive, also shows generalization across scenarios. However, its more rigid control behaviors–such as aggressive pre-cooling–can result in inefficiencies under certain conditions (e.g., when occupancy begins late on hot days). Despite this, it generally performs on par with G36 in terms of comfort and CO_2 levels.

7. Conclusion

This study investigated the use of Deep Reinforcement Learning for optimizing HVAC operations in a multi-zone educational building, using a high-fidelity Modelica simulation environment. Two low-level DRL controllers–Zone-Aware and Zone-Integrated–were developed and evaluated against the ASHRAE G36 control standard. The controllers directly act on full low-level Air Handling Unit (AHU) components, such as fan speed, damper positions, and coil valves, while leaving zone-level VAV control unaltered.

The proposed DRL strategies were assessed under both heating and cooling conditions, as well as across different occupancy profiles. Performance was evaluated in terms of energy consumption, indoor temperature control, and CO_2 levels.

The results demonstrate that DRL-based controllers effectively manage HVAC operations under trained and previously unseen conditions, showcasing their ability to optimize energy use while maintaining indoor temperature tracking and CO_2 concentration. The Zone-Aware controller emerged as the most effective, highlighting the benefit of zone-specific feedback. However, the Zone-Integrated DRL also proved to be a viable alternative, with similar performance in most scenarios and the added advantage of being potentially more scalable for large zone buildings.

While the heating season presented some limitations due to the constraints of the VAVs' reheating system, the cooling season revealed the full potential of DRL controllers–particularly the Zone-Aware strategy, which adapted effectively to varying occupancy and load conditions. The Zone-Integrated DRL, though slightly less responsive, also performed competitively, although its rigid control behavior under certain conditions suggests areas for refinement.

Overall, the findings suggest that DRL-based controllers offer a promising approach to improve HVAC efficiency, with significant potential for adaptation across dynamic building environments. The success of the Zone-Integrated controller further indicates that less granular control strategies may offer practical, scalable solutions for large or resource-constrained applications. The study highlights the potential of DRL as a technology for energy-efficient building operation, capable of adapting to real-time conditions without the need for manual intervention.

8. Limitations and future work

While the results demonstrate the strong performance and generalization capabilities of DRL-based HVAC controllers, some limitations should be acknowledged.

First, the evaluation primarily focuses on changes in occupancy schedules, while other critical boundary conditions–such as outdoor temperature, solar radiation, and outdoor CO_2 levels–remain constant across training and testing. Yet, these factors play a crucial role in real-world HVAC performance and may challenge the adaptability of learned policies. Although preliminary tests with alternative weather datasets have shown that DRL performance remains promising, future work will systematically extend the evaluation to more diverse climatic profiles, as well as dynamic external CO_2 levels, to more comprehensively demonstrate the adaptability of the proposed DRL strategies.

Second, the study is conducted within a relatively homogeneous zone context, using classroom spaces that usually have similar thermal loads and occupancy patterns. While this provides a simplified yet consistent context to evaluate the core behavior of the proposed DRL strategies, it may also limit the exploration of more complex dynamics–particularly for the Zone-Integrated DRL, which relies on aggregated, global observations. To address these limitations, future experiments should consider case studies including a wider variety of functional areas, such as offices, conference rooms, corridors, and circulation spaces, which are characterized by more heterogeneous thermal and occupancy profiles. Such diversity could significantly influence the balance between

comfort and energy trade-offs in DRL control of a VAV multi-zone AHU system. Therefore, extending the evaluation to mixed-use buildings with diverse load characteristics will allow a more rigorous assessment of the proposed DRL framework.

Future work could include a comparative study evaluating the low-level DRL approach presented in this work against DRL strategies that operate at a higher level by selecting setpoints for supply air temperature and duct static pressure. Such a comparison would help quantify the advantages and trade-offs between direct low-level control and supervisory control approaches.

Further investigation could also focus on applying transfer learning techniques to enhance the adaptability of DRL controllers under varying boundary conditions [46–49]. In addition, extending these methods to scenarios with more diverse thermal demand profiles across zones would provide deeper insights into policy generalization and control accuracy–particularly for zone-integrated strategies.

Finally, a promising direction lies in the interpretation and extraction of human-readable rules from trained DRL policies [39,50–52]. While qualitative patterns in DRL control policies can be observed, translating them into quantitative rules with explicit thresholds (e.g., for supply air temperature or fan speed) is not straightforward. Automated rule-extraction techniques can overcome this limitation by systematically identifying the most relevant features, their critical ranges, and the corresponding control actions, thus enabling the derivation of interpretable control rules.

Beyond these technical aspects, it is essential to acknowledge that real-world buildings present additional sources of uncertainty and practical constraints not captured by simulation models. Future work should therefore focus on field experiments to validate the applicability and robustness of low-level DRL-based control strategies. Particular attention should be paid to discrepancies between simulated and actual AHU configurations, variations in operating conditions, deployment and maintenance costs, and challenges related to data communication and command response. Considering these factors will be critical to rigorously assess the feasibility of deploying DRL-based HVAC controllers in real building environments.

CRediT authorship contribution statement

Sabrina Savino: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Giuseppe Razzano:** Writing – original draft, Software, Methodology, Investigation, Data curation; **Michele Pagone:** Writing – review & editing, Validation; **Carlo Novara:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization; **Alfonso Capozzoli:** Writing – review & editing, Validation, Supervision, Project administration, Methodology.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The work of Sabrina Savino is part of the project PNRR-NGEU- CUP: E14D23001820006 which has received funding from the MUR - DM 118/2023. The work of Alfonso Capozzoli and Giuseppe Razzano was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)-MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3-D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] I.E. Agency, Buildings - Topics - IEA, 2025. Accessed: 2025-02-05, https://www.iea.org/energy-system/buildings.
[2] E.K. Simpeh, J.P.G. Pillay, R. Ndihokubwayo, D.J. Nalumu, Improving energy efficiency of HVAC systems in buildings: a review of best practices, 2022. https://doi.org/10.1108/IJBPA-02-2021-0019
[3] P. Wargocki, J. Sundell, W. Bischof, G. Brundrett, P.O. Fanger, F. Gyntelberg, S.O. Hanssen, P. Harrison, A.C. Pickering, O. Seppänen, P. Wouters, Ventilation and health in non-industrial indoor environments: report from a european multidisciplinary scientific consensus meeting (EUROVEN), Indoor Air 12 (2) (2002) 113–128. https://api.semanticscholar.org/CorpusID:9440736t.
[4] O. Seppänen, W.J. Fisk, M.J. Mendell, Association of ventilation rates and CO2 concentrations with health and other responses in commercial and institutional buildings, Indoor Air 9 (4) (1999) 226–252. https://api.semanticscholar.org/CorpusID:14264134.
[5] M. Eydner, B. Toufek, T. Henzler, K. Stergiaropoulos, Investigation of a multizone building with HVAC system using a coupled thermal and airflow model, E3S Web Conf. (2019). https://api.semanticscholar.org/CorpusID:202191401.
[6] L. Yang, P. Xu, Y. Li, Nonlinear dynamic analysis of natural ventilation in a two-zone building: Part A–theoretical analysis, HVAC&R Res. 12 (2006) 231–255. https://api.semanticscholar.org/CorpusID:121140965.
[7] J.P. Koeln, B.D. Keating, A.G. Alleyne, C.R. Price, B.P. Rasmussen, Multi-Zone Temperature Modeling and Control, 2018. https://api.semanticscholar.org/CorpusID:139141664.
[8] ASHRAE, ASHRAE Guideline 36–2021: High-Performance Sequences of Operation for HVAC Systems, American Society of Heating, Refrigerating, and Air-Conditioning Engineers, 2021.
[9] K. Zhang, D. Blum, H. Cheng, G. Paliaga, M. Wetter, J. Granderson, Estimating ASHRAE guideline 36 energy savings for multi-zone variable air volume systems using spawn of energyplus, J. Build. Perform. Simul. 15 (2022) 215–236. https://doi.org/10.1080/19401493.2021.2021286
[10] E.F. Camacho, C. Bordons, Model Predictive Control, Springer, London, UK, 2013.
[11] D. Kim, J.E. Braun, Development, implementation and performance of a model predictive controller for packaged air conditioners in small and medium-sized commercial building applications, Energy Build. 178 (2018) 49–60. https://doi.org/10.1016/j.enbuild.2018.08.019
[12] M. Tanaskovic, D. Sturzenegger, R. Smith, M. Morari, Robust adaptive model predictive building climate control, in: IFAC-PapersOnLine, 50, Elsevier B.V., 2017, pp. 1871–1876. https://doi.org/10.1016/j.ifacol.2017.08.257
[13] S.w. Ham, D. Kim, T. Barham, K. Ramseyer, The first field application of a low-cost MPC for grid-interactive K-12 schools: lessons-learned and savings assessment, Energy Build. 296 (2023). https://doi.org/10.1016/j.enbuild.2023.113351
[14] H. Zhang, S. Seal, D. Wu, F. Bouffard, B. Boulet, Building energy management with reinforcement learning and model predictive control: a survey, IEEE Access 10 (2022) 27853–27862. https://doi.org/10.1109/ACCESS.2022.3156581
[15] M. Mork, N. Materzok, A. Xhonneux, D. Müller, Nonlinear hybrid model predictive control for building energy systems, Energy Build. 270 (2022) 112298. https://doi.org/10.1016/j.enbuild.2022.112298
[16] S. Sharma, A. Srinivas, B. Ravindran, Deep reinforcement learning, 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (2023) 1–7. https://api.semanticscholar.org/CorpusID:15294098t.
[17] M. Wetter, Co-simulation of building energy and control systems with the building controls virtual test bed, J. Build. Perform. Simul. 4 (3) (2011) 185–203. https://doi.org/10.1080/19401493.2010.518631
[18] K. Kurte, J. Munk, O. Kotevska, K. Amasyali, R. Smith, E. McKee, Y. Du, B. Cui, T. Kuruganti, H. Zandi, Evaluating the adaptability of reinforcement learning based HVAC control for residential houses, Sustainability 12 (18) (2020). https://doi.org/10.3390/su12187727
[19] Q. Alfalouji, T. Schranz, B. Falay, S. Wilfling, J. Exenberger, T. Mattausch, C. Gomes, G. Schweiger, Co-simulation for buildings and smart energy systems – A taxonomic review, 2023. https://doi.org/10.1016/j.simpat.2023.102770
[20] T. Blockwitz, M. Otter, J. Akesson, M. Arnold, C. Clauss, H. Elmqvist, M. Friedrich, A. Junghanns, J. Mauss, D. Neumerkel, H. Olsson, A. Viel, Functional Mockup Interface 2.0: The Standard for Tool Independent Exchange of Simulation Models, 76, Linköping University Electronic Press, 2012, pp. 173–184. https://doi.org/10.3384/ecp12076173
[21] Y. Du, F. Li, J. Munk, K. Kurte, O. Kotevska, K. Amasyali, H. Zandi, Multi-task deep reinforcement learning for intelligent multi-zone residential HVAC control, Electr. Power Syst. Res. 192 (2021). https://doi.org/10.1016/j.epsr.2020.106959
[22] Y. Du, H. Zandi, O. Kotevska, K. Kurte, J. Munk, K. Amasyali, E. Mckee, F. Li, Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning, Appl. Energy 281 (2021). https://doi.org/10.1016/j.apenergy.2020.116117
[23] G. Gao, J. Li, Y. Wen, Deepcomfort: energy-efficient thermal comfort control in buildings via reinforcement learning, IEEE Internet Things J. 7 (2020) 8472–8484. https://doi.org/10.1109/JIOT.2020.2992117

[24] X. Deng, Y. Zhang, H. Qi, Towards optimal HVAC control in non-stationary building environments combining active change detection and deep reinforcement learning, Build. Environ. 211 (2022). https://doi.org/10.1016/j.buildenv.2021.108680
[25] X. Lu, Y. Fu, Z. O'Neill, Benchmarking high performance HVAC rule-based controls with advanced intelligent controllers: a case study in a multi-zone system in modelica, Energy Build. 284 (2023). https://doi.org/10.1016/j.enbuild.2023.112854
[26] C. Cui, J. Xue, Energy and comfort aware operation of multi-zone HVAC system through preference-inspired deep reinforcement learning, Energy 292 (2024). https://doi.org/10.1016/j.energy.2024.130505
[27] E. Zanetti, D. Kim, D. Blum, R. Scoccia, M. Aprile, Performance comparison of quadratic, nonlinear, and mixed integer nonlinear MPC formulations and solvers on an air source heat pump hydronic floor heating system, J. Build. Perform. Simul. 16 (2023) 144–162. https://doi.org/10.1080/19401493.2022.2120631
[28] K.U. Ahn, C.S. Park, Application of deep Q-networks for model-free optimal control balancing between different HVAC systems, Sci. Technol. Built Environ. 26 (2020) 61–74. https://doi.org/10.1080/23744731.2019.1680234
[29] M. Biemann, F. Scheller, X. Liu, L. Huang, Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control, Appl. Energy 298 (2021). https://doi.org/10.1016/j.apenergy.2021.117164
[30] F. Guo, S.w. Ham, D. Kim, H.J. Moon, Deep reinforcement learning control for co-optimizing energy consumption, thermal comfort, and indoor air quality in an office building, Appl. Energy 377 (2025). https://doi.org/10.1016/j.apenergy.2024.124467
[31] Y. Liu, Y. Song, C. Cui, Towards smart control and energy efficiency for multi-zone ventilation systems via an imitation-interaction learning method in energy-aware buildings, Energy 314 (2025). https://doi.org/10.1016/j.energy.2024.134220
[32] P. Fritzson, P. Aronsson, H. Lundvall, K. Nyström, A. Pop, L. Saldamli, D. Broman, The openmodelica modeling, simulation, and development environmentSimulation News Europe (2005).
[33] U. S. Department of Energy, EnergyPlus™ Energy Simulation Software, 2023, https://energyplus.net. Version 23.1.0.
[34] D. Blum, J. Arroyo, S. Huang, J. Drgoňa, F. Jorissen, H.T. Walnum, Y. Chen, K. Benne, D. Vrabie, M. Wetter, L. Helsen, Building optimization testing framework (BOPTEST) for simulation-based benchmarking of control strategies in buildings, J. Build. Perform. Simul. 14 (2021) 586–610. https://doi.org/10.1080/19401493.2021.1986574
[35] C. Blad, S. Bøgh, C.S. Kallesøe, Data-driven offline reinforcement learning for HVAC-systems, Energy 261 (2022). https://doi.org/10.1016/j.energy.2022.125290
[36] Y. Zhang, Q. Zhao, Energy saving algorithm of HVAC system based on deep reinforcement learning with modelica model, Proceedings of the 2022 41st Chinese Control Conference (CCC) (2022).
[37] Modelica Association, Modelica - A Unified Object-Oriented Language for Systems Modeling, 2021. https://www.modelica.org.
[38] S. Brandi, A. Pizza, G. Buscemi, G. Raz-Zano, A. Capozzoli, Enhancing Energy Efficiency and Flexibility in Educa-tional Buildings through a Deep Reinforcement Learn-ing-Based Controller for Rooftop Units, Technical Report, (2024)
[39] G. Razzano, S. Brandi, M.S. Piscitelli, A. Capozzoli, Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal management of heating, ventilation, and air conditioning (HVAC) systems in multizone buildings, Appl. Energy 381 (2025). https://doi.org/10.1016/j.apenergy.2024.125046
[40] M. Wetter, W. Zuo, T.S. Nouidui, X. Pang, Modelica buildings library, J. Build. Perform. Simul. 7 (4) (2014) 253–270. https://doi.org/10.1080/19401493.2013.765506
[41] C. Andersson, J. Åkesson, C. Führer, PyFMI: A Python Package for Simulation of Coupled Dynamic Models with the Functional Mock-up Interface, 2016. https://api.semanticscholar.org/CorpusID:218002023.
[42] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, W. Zaremba, OpenAI gym (2016). http://arxiv.org/abs/1606.01540.
[43] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, N. Dormann, Stable-baselines3: reliable reinforcement learning implementations, J. Mach. Learn. Res. 22 (268) (2021) 1–8. http://jmlr.org/papers/v22/20-1364.html.
[44] OneBuilding.org, Climate Data for Building Simulation, 2023. Accessed: 2024-04-10, https://climate.onebuilding.org/.
[45] Energy Performance of Buildings-Ventilation for Buildings, 2019.
[46] K. Kadamala, D. Chambers, E. Barrett, Enhancing HVAC control systems through transfer learning with deep reinforcement learning agents, Smart Energy 13 (2024). https://doi.org/10.1016/j.segy.2024.100131
[47] M. Esrafilian-Najafabadi, F. Haghighat, Transfer learning for occupancy-based HVAC control: a data-driven approach using unsupervised learning of occupancy profiles and deep reinforcement learning, Energy Build. 300 (2023) 113637. https://doi.org/10.1016/j.enbuild.2023.113637
[48] D. Coraci, A. Silvestri, G. Razzano, D. Fop, S. Brandi, E. Borkowski, T. Hong, A. Schlueter, A. Capozzoli, A scalable approach for real-world implementation of deep reinforcement learning controllers in buildings based on online transfer learning: the hilo case study, Energy Build. 329 (2025) 115254. https://doi.org/10.1016/j.enbuild.2024.115254
[49] P. Lissa, M. Schukat, E. Barrett, Transfer learning applied to reinforcement learning-Based HVAC control, SN Comput. Sci. 1 (3) (2020) 127. https://doi.org/10.1007/s42979-020-00146-7
[50] Y. Choi, X. Lu, Z. O'Neill, F. Feng, T. Yang, Optimization-informed rule extraction for HVAC system: a case study of dedicated outdoor air system control in a mixed-humid climate zone, Energy Build. 295 (2023). https://doi.org/10.1016/j.enbuild.2023.113295
[51] S. Cho, C.S.P. and, Rule reduction for control of a building cooling system using explainable AI, J. Build. Perform. Simul. 15 (6) (2022) 832–847 https://doi.org/10.1080/19401493.2022.2103586
[52] B. Gunay, M. Ouf, G. Newsham, W. O'Brien, Building performance optimization for operational rule extraction, in: Proceedings of Building Simulation 2019: 16th Conference of IBPSA, 16 of *Building Simulation*, IBPSA, Rome, Italy, 2019, pp. 2819–2826. https://doi.org/10.26868/25222708.2019.210271